Um Modelo de Risco Proporcional Dependente do Tempo

Daniela Ribeiro Martins Parreira

Co-orientadora: Profa. Dra. Vera Lucia Tomazella
Orientador: Prof. Dr. Francisco Louzada-Neto

Dissertação apresentada ao Departamento de Estatística da Universidade Federal de São Carlos - DEs/UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística.

São Carlos
Junho de 2007

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

P436mr

Parreira, Daniela Ribeiro Martins.
Um modelo de risco proporcional dependente do tempo / Daniela Ribeiro Martins Parreira. -- São Carlos : UFSCar, 2008.
50 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2007.

1. Estatística matemática. 2. Modelagem. 3. Modelo de Cox. 4. Riscos proporcionais. 5. Inferência bayesiana. 6. Inferência clássica. I. Título.

CDD: 519.54 (20ª)

Dedico esta tese às pessoas que se fizeram presentes em minha vida.

Meus pais, Osvaldo e Delizete, inspiração desde os primeiros e vacilantes passos até os mais ousados. Incentivaram-me e sempre acreditaram em mim. Efusiva homenagem e...uma gratidão sem fim.

Meu irmão, Osvaldo Jr., e Fernanda. Bárbara e Laura, adoráveis sobrinhas, esperança dos mais doces sonhos para toda a família.

Leandro, que, com sua constante presença, afetuoso convívio e colaboração, tornou suave este empreendimento.

Professor Doutor José Galvão Leite. Meu especial reconhecimento pelo seu competente magistério, atenção magnânima, gestos bondosos, motivo de estímulo em todas as fases de investigação.

Professora Doutora Vera Lúcia Tomazella e Professor Doutor Francisco Louzada-Neto, meus orientadores. A eficiência com que acompanharam o desenvolvimento das pesquisas, seus esclarecimentos e indicações concorreram, imprescindivelmente, para a organização geral desta reflexão. Minha admiração e respeito.

Professora Doutora Tereza Cristina e Professor Doutor Benedito Benze. Grata pela generosidade, valioso concurso e magistrais sugestões, ainda, no período da qualificação.

Aos meus tios Jonas e Waldete por me terem como uma filha, e assim, dedicar a mim um carinho único. Agradeço tudo o que têm feito por mim desde pequena.

Amigos. Àqueles que me apoiaram o débito é imenso e o espaço diminuto para expressar toda a minha alegria por esta realização.

Resumo

A análise de sobrevivência tem por objetivo estudar dados de experimento em que a variável resposta é o tempo até a ocorrência de um evento de interesse. Vários autores têm preferido modelar dados de sobrevivência na presença de covariáveis por meio da função de risco, fato este relacionado à sua interpretação. Ela descreve como a probabilidade instantânea de falha se modifica com o passar do tempo. Nesse contexto, um dos modelos mais utilizados é o modelo de Cox (Cox, 1972), onde a suposição básica para o seu uso é que as taxas de falhas sejam proporcionais. O modelo de riscos proporcionais de Cox é bastante flexível e extensivamente usado em análise de sobrevivência. Ele pode ser facilmente estendido para incorporar, por exemplo, o efeito de covariáveis dependentes do tempo. Neste estudo, propõe-se um modelo de risco proporcional, que incorpora um parâmetro dependente do tempo, denominado modelo de risco proporcional dependente do tempo. Uma análise clássica baseada nas propriedades assintóticas dos estimadores de máxima verossimilhança dos parâmetros envolvidos é desenvolvida, bem como um estudo de simulação via técnicas de reamostragem para estimação intervalar e testes de hipóteses dos parâmetros do modelo. É estudado o custo de estimar o efeito da covariável quando o parâmetro que mede o efeito do tempo é considerado na modelagem. E, finalizando, apresentamos uma abordagem do ponto de vista Bayesiano.

Palavras-chave: Análise de Sobrevivência, Funções de Risco, Covariáveis, Modelo de Risco Proporcional de Cox, Modelo de Risco Proporcional Dependente do Tempo.

Abstract

Survival data analysis models is used to study experimental data where, normally, the variable "answer"is the time passed until an event of interest. Many authors do prefer modeling survival data, in the presence of co-variables, by using a hazard function – which is related with its interpretation. The Cox model (1972) – most commonly used by the authors – is applicable when the fail rates are proportional. This model is very flexible and used in the survival analysis. It can be easily extended to, for example, incorporate the time-dependent co-variables. In the present work we propose a proportional risk model which incorporates a time-dependent parameter named "time-dependent proportional risk model".

Key-words: Survival analysis, Risk functions, co-variables, Cox's proportional Risk models, Time-dependent proportional risk model.

Sumário

1 Introdução **1**

1.1 Análise de Sobrevivência 1

1.1.1 A Presença de Censuras 3

1.1.2 A Presença de Variáveis Explicativas 4

1.1.3 A Função de Verossimilhança 5

1.2 Modelagem Via Função de Risco 5

1.2.1 O Modelo de Riscos Proporcionais 6

1.2.2 O Modelo de Falha Acelerada 7

1.2.3 O Modelo Híbrido 7

1.2.4 O Modelo Híbrido Estendido 7

1.3 Conteúdo da Dissertação 8

2 Modelo de Risco Proporcional Dependente do Tempo **9**

2.1 Introdução 9

2.2 Formulação do Modelo 9

2.3 Propriedades do Modelo 11

2.4 Procedimento de Estimação Pontual 13

2.5 Procedimento de Estimação Intervalar 15

2.6 Aplicação 15

2.7 Estudo Numérico 19

2.7.1 Cálculo da Probabilidade de Cobertura 19

2.8 Considerações Finais 21

3 Método de Simulação *Bootstrap* 22

3.1 Intervalo de Confiança via Método *Bootstrap* 23

3.1.1 Aplicação 24

3.1.2 Probabilidade de Cobertura dos Intervalos de Confiança *Bootstrap* . 24

3.2 Considerações Finais 25

4 O Custo de Estimar β na presença de α 26

4.1 Introdução 26

4.2 Definindo o Custo de Estimar β 27

4.3 Estudo Numérico 28

4.4 Conclusões 29

5 Abordagem Bayesiana 30

5.1 Introdução 30

5.2 Metodologia 30

5.2.1 Cadeis de Markov Monte Carlo (MCMC) 31

5.2.2 Metropolis-Hastings 32

5.2.3 Diagnósticos de Convergência 33

5.3 Aplicação 34

6 Conclusões e Perspectivas Futuras 39

Apêndices 40

Referências Bibliográficas 41

Apêndices 45

A O Modelo de Gompertz 46

A.1 A Função de Verossimilhança de Gompertz 47

B O Modelo de Riscos Proporcionais de Cox 48

B.1 A Função de Verossimilhança Parcial de Cox 49

Capítulo 1

Introdução

1.1 Análise de Sobrevivência

A uma coleção de procedimentos estatísticos para a análise de dados relacionados com o tempo até a ocorrência de um determinado evento de interesse, a partir de um tempo inicial preestabelecido, chama-se de análise de sobrevivência ou confiabilidade. Geralmente, a aplicação da análise de sobrevivência se faz presente na área médica, enquanto que a análise de confiabilidade refere-se à pesquisa industrial.

Algumas peculiaridades podem ser encontradas nos dados de sobrevivência, como a presença de censuras e de covariáveis, a quantidade de causa de falha e o número de eventos recorrentes. A censura indica se o valor do tempo de sobrevivência de determinado indivíduo foi observado ou não. A censura impede o uso de procedimentos estatísticos convencionais tais como a análise de regressão e a análise de planejamento de experimentos, uma vez que tais procedimentos não permitem que as informações contidas nos dados censurados sejam incorporadas. Sendo assim, há necessidade de uma metodologia apropriada para esse tipo de problema.

A variável de interesse, tempo de vida (sobrevivência) ou até a falha, é estritamente positiva e, geralmente, medida em escala contínua.

Em análise de sobrevivência, as unidades em estudo são, usualmente, indivíduos. Essas unidades compõem os dados de sobrevivência, formados essencialmente pelos tempos de

vida até a ocorrência de algum evento de interesse. Os tempos são aqui chamados de tempos de sobrevivência, podendo, algumas vezes, ser referidos como tempos de falhas ou simplesmente tempos de vida.

Os dados de sobrevivência incorporam tanto os tempos de sobrevivência como um conjunto de variáveis observáveis que podem estar relacionadas com os mesmos. Essas variáveis são conhecidas por covariáveis, variáveis explicativas ou variáveis explanatórias. Quando os tempos de sobrevivência estão relacionados com covariáveis, diz-se que a população das suas unidades é heterogênea, caso contrário, a sua população é dita homogênea. As covariáveis são geralmente medidas uma única vez ao longo do tempo. Contudo, podem-se encontrar, em alguns casos, dados de sobrevivência em que as covariá- veis estão em função do tempo, pelo fato de seus valores serem modificados ao longo do período de observação.

O comportamento da variável aleatória contínua e não negativa tempo de sobrevivência, $T \geq 0$, pode ser expresso através de várias funções, matematicamente, equivalentes, tal que, se uma delas é especificada, as outras podem ser derivadas. Entre elas tem-se: a função de densidade de probabilidade, $f(t)$, a função de sobrevivência, $S(t)$, e a função de risco, $h(t)$.

A função densidade de probabilidade (f.d.p.) é definida como o limite da probabilidade de um indivíduo falhar no intervalo de tempo $[t, t+dt)$ por unidade de tempo, e é expressa por (Lee, 1992 pg. 11),

$$f(t) = \lim_{dt \to 0} \frac{P(t \leq T < t+dt)}{dt}, \tag{1.1}$$

onde $f(t) \geq 0$ para todo t, e tem a área abaixo da curva igual a 1.

A função de sobrevivência é definida como sendo a probabilidade de um indivíduo sobreviver mais que determinado tempo, t, dada por (Lawless, 1982 pg.8),

$$S(t) = P(T \geq t) = 1 - F_T(t). \tag{1.2}$$

Uma consequência desta distribuição é a respectiva função de distribuição acumulada $F(t) = 1 - S(t), \forall\, t \in \Im = [0, \infty)$. A função de sobrevivência é também conhecida como taxa de sobrevivência acumulada.

A função de risco é definida pelo limite da probabilidade de um indivíduo falhar no in-

tervalo de tempo $[t, t+dt)$para dt tendendo a zero, dado que ele mesmo tenha sobrevivido até o instante t, e é expressa por (Cox & Oakes, 1984 pg. 14),

$$h(t) = \lim_{dt \to 0^+} \frac{P(t \leq T < t + dt | T \geq t)}{dt}. \tag{1.3}$$

Devido à sua interpretação a função de risco tem sido preferida por muitos autores para descrever o comportamento do tempo de sobrevivência. Ela descreve como a probabilidade instantânea de falha se modifica com o passar do tempo, e ainda, através dela podemos caracterizar classes especiais de distribuições de tempo de sobrevivência.

A função de risco $h(t)$ também pode ser definida em termos (1.1) e (1.2) por meio da expressão,

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dx} \log S(x) \tag{1.4}$$

e

$$S(t) = \exp\left[-\int_0^t h(u)\, du\right]. \tag{1.5}$$

Assim, temos de (1.4) e (1.5) que a fdp pode ser reescrita como,

$$f(t) = h(t)\, S(t) = h(t) \exp\left[-\int_0^t h(u)\, du\right]. \tag{1.6}$$

1.1.1 A Presença de Censuras

Um complicador presente nos dados de sobrevivência é o fato da variável de interesse, tempo de sobrevivência, não ser medida instantânea e independentemente do tamanho da resposta. Valores grandes desta variável necessitam de mais tempo e persistência para ser observados. Em situações extremas, esse fato comprometerá a observação do valor da variável para alguns indivíduos, uma vez que o evento de interesse pode não ocorrer até o final do tempo de estudo. Também, o paciente poderá abandonar a investigação antes mesmo da observação do evento de interesse ou falecer devido a outras causas, que não a causa em evidência. Isto pode, inclusive, acontecer antes mesmo do tempo final de recrutamento. Da mesma maneira, esse tipo de problema é apresentado em dados industriais, quando somente uma proporção de componentes em questão falha até o final do estudo.

Essas informações devem estar incorporadas na análise dos dados, sendo assim, existe a necessidade da introdução de uma variável extra, indicando se o valor do tempo de sobrevivência de um determinado indivíduo foi ou não observado. Essa variável é conhecida, na literatura de análise de sobrevivência e confiabilidade, como variável indicadora de censuras, ou simplesmente censuras.

Sem a presença de censuras, a análise se reduz à utilização das técnicas estatísticas clássicas, como a análise de regressão e métodos de planejamento de experimentos.

As censuras são classificadas, de um modo geral, como:

Censura tipo I: Os tempos de sobrevivência são maiores que o tempo final do experimento a partir de um tempo final do estudo preestabelecido.

Censura tipo II: O estudo termina depois que um número preestabelecido de falhas ocorra.

Censuras aleatórias: quando um indivíduo é retirado do estudo, sem ter ocorrido falha, ou pela ocorrência de um evento diferente daquele de interesse.

Outros tipos de censuras não são considerados aqui mas podem ser encontrados em Lawless (1982) e Collett (1994).

1.1.2 A Presença de Variáveis Explicativas

Além do tempo de sobrevivência e da variável indicadora de censuras, podem também ser observadas, nos dados, variáveis que representam tanto a heterogeneidade existente na população, tais como idade, sexo, entre outras, como também possíveis tratamentos aos quais os indivíduos são submetidos. Estas variáveis são conhecidas como variáveis explicativas ou covariáveis.

Muitas vezes, o objetivo da análise de sobrevivência está centrado na relação entre o tempo de sobrevivência e algumas variáveis explicativas de interesse. Do ponto de vista estatístico, tem-se as variáveis tempo de sobrevivência, variável indicadora de censuras, e um vetor de variáveis explicativas disponíveis para a análise.

Uma complicação adicional que também pode ocorrer na análise de sobrevivência e confiabilidade é encontrar variáveis explicativas que dependem do tempo, ou seja, os valores das covariáveis, no final do experimento, podem não ser os mesmos que no seu início. Por exemplo, pode-se ter um experimento em que a dose de um determinado

medicamento é modificada ao longo do experimento, uma vez que, por exemplo, o paciente apresente efeitos colaterais à droga.

1.1.3 A Função de Verossimilhança

Considerando que a presença de tempos censurados é comum no contexto de análise de sobrevivência, suponha que alguns dos tempos observados sejam censurados à direita. Suponha também, um esquema de censuras onde estas ocorrem em tempos diferentes, de indivíduo para indivíduo, e que tal esquema sempre possa ser considerado estatisticamente independente do mecanismo que causa a morte do indivíduo (ou ocorrência da característica em observação), ou seja, um esquema de censuras independentes dos tempos de sobrevivência e das covariadas.Sejam Y_i o tempo exato de ocorrência de determinada falha (característica) de interesse, L_i tempo pré-fixado de estudo,γ_i variável indicadora de censura e T_i tempo de sobrevivência.

Desta forma, os tempos de sobrevivência T_i e as variáveis indicadoras de censura γ_i $i = 1, ..., n$ são definidas por:

$$T_i = \begin{cases} Y_i, \text{ se } Y_i \leq L_i \\ L_i, \quad \text{se } Y_i > L_i \end{cases} \quad \text{e} \quad \gamma_i = \begin{cases} 1, \text{ se } Y_i \leq L_i \\ 0, \quad \text{se } Y_i > L_i \end{cases}.$$

Considerando (1.4) e (1.5) e a presença de censuras, de acordo com o esquema descrito acima, a função de verossimilhança é dada por

$$L(\mathbf{t}) = \prod_{i=1}^{n} f(t_i)^{\gamma_i} S(L_i)^{1-\gamma_i} = \prod_{i=1}^{n} h(t_i|x_i)^{\gamma_i} S(t_i|x_i) \tag{1.7}$$

1.2 Modelagem Via Função de Risco

Modelos de risco têm sido amplamente utilizados para modelar dados de sobrevivência. Um dos mais empregados é o modelo de Cox (1972) que foi o primeiro modelo proposto para modelar dados de sobrevivência na presença de covariáveis. Apresentamos aqui a evolução dos modelos de risco, à partir do modelo de Cox. Nota-se claramente que, conforme evoluíram, suas estruturas se tornaram mais flexíveis.

Isto permite o estabelecimento de modelos mais apropriados para determinados conjuntos de dados e também facilita o estudo da má especificação dos mesmos.

Considerando a presença de covariáveis, a função de risco pode ser escrita da forma,

$$h\left(t|\mathbf{x}\right) = u\left(t, \boldsymbol{\alpha}^{'}\mathbf{x}\right), \tag{1.8}$$

onde $u\left(.\right)$ é uma função positiva, igual a 1 quando seu argumento é 0, $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_p)$ representa o vetor de covariáveis e $\boldsymbol{\alpha}^{'} = (\alpha_1, ..., \alpha_p)$ de coeficientes desconhecidos de regressão.

1.2.1 O Modelo de Riscos Proporcionais

Considere-se que a função de risco (1.8) pode ser fatorada em dois termos, um que represente o efeito das covariáveis e outro o do tempo. Seja $h\left(t|\mathbf{x}\right)$ a função de risco no tempo t para um indivíduo com vetor de covariáveis $\mathbf{x}$.

O modelo de riscos proporcionais proposto por Cox (1972) (PH), é dado por,

$$h\left(t|\mathbf{x}\right) = g\left(\boldsymbol{\alpha}^{'}\mathbf{x}\right) h_0\left(t\right), \tag{1.9}$$

onde $g(\cdot)$ é uma função positiva, que assume o valor 1 quando seu argumento é igual a zero, $h_0\left(\cdot\right)$ representa a função de risco básica para uma unidade quando $\mathbf{x} = \mathbf{0}$ e $\boldsymbol{\alpha}^{'}$ é o vetor de coeficientes a serem estimados. Várias formas funcionais podem ser empregadas para $g\left(\cdot\right)$, entretanto a candidata natural, que será aqui tratada, é a função $\exp\left(\cdot\right)$.

Este modelo assume que o vetor de covariáveis $\mathbf{x}$ tem um efeito multiplicativo na função de risco. Isto implica que a sua estrutura impõe proporcionalidade entre funções de risco de diferentes níveis de covariáveis, não permitindo que elas se cruzem e dependam do tempo t.

A grande limitação do modelo de riscos proporcionais é a suposição de proporcionalidade entre as funções de risco em diferentes níveis das covariáveis, uma vez que na prática não é difícil encontrar-se funções de riscos não proporcionais.

1.2.2 O Modelo de Falha Acelerada

Em situações em que o modelo de Cox não é adequado, tem-se o modelo de falha acelerada (AFT) dado por Kalbfleish e Prentice (1980),

$$h\left(t|\mathbf{x}\right) = g\left(\boldsymbol{\alpha}'\mathbf{x}\right) h_0\left(g\left(\boldsymbol{\alpha}'\mathbf{x}\right) t\right). \tag{1.10}$$

Uma vantagem deste modelo é permitir o cruzamento de funções de risco, uma vez que $\mathbf{x}$ tem um efeito multiplicativo não somente na função de risco, mas também em t. Desta forma, $\mathbf{x}$ afeta o tempo de sobrevivência causando deformações na escala de tempo.

Do ponto de vista interpretativo, os modelos (1.9) e (1.10) compreendem famílias distintas, devendo então, ser tratados como tal (Louzada Neto, 1997).

1.2.3 O Modelo Híbrido

Com o intuito de acomodar ambos os modelos em uma família, Ciampi e Etezade-Amoli (1985) propuseram um modelo Híbrido (PH/AFT) dado por,

$$h\left(t|\mathbf{x}\right) = g\left(\boldsymbol{\alpha}_1'\mathbf{x}\right) h_0\left(g\left(\boldsymbol{\alpha}_2'\mathbf{x}\right) t\right), \tag{1.11}$$

onde $\boldsymbol{\alpha}_1$ e $\boldsymbol{\alpha}_2$ são vetores de coeficientes. Para $\alpha_2 = 0$, obtem-se o modelo de riscos proporcionais, e, para $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2$, tem-se o modelo de falha acelerada.

Estes modelos são muito utilizados na análise de dados de sobrevivência. Entretanto, os mesmos não possuem capacidade de acomodar uma possível "heterocedasticidade", ou seja, variâncias diferentes sobre indivíduos de diferentes níveis de covariáveis.

1.2.4 O Modelo Híbrido Estendido

Considerando que os modelos (1.9), (1.10) e (1.11) não possuem a capacidade de acomodar a presença de "heterocedasticidade", Louzada Neto (1997) propôs um modelo de risco híbrido estendido, permitindo que o parâmetro de forma dependa de covariáveis por meio de uma função chamada "heteroscedástica". Este modelo é dado por,

$$h\left(t|\mathbf{x}\right) = g\left(\gamma' x\right) \left[g\left(\boldsymbol{\alpha}_1'\mathbf{x}\right) h_0\left(g\left(\boldsymbol{\alpha}_2'\mathbf{x}\right) t\right)\right]^{g\left(\gamma'\mathbf{x}\right)}, \tag{1.12}$$

onde $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$ e $\boldsymbol{\gamma}$ são vetores de coeficientes de regressão desconhecidos, e a função $g\left(\boldsymbol{\gamma}'\mathbf{x}\right)$ que é igual a $\exp\left(\boldsymbol{\gamma}'\mathbf{x}\right)$ descreve a relação "heterocedástica" entre o parâmetro de forma e as covariáveis.

As diferenças entre as funções de risco e de sobrevivência, dos modelos (1.9), (1.10) e (1.11) são mostradas na Tabela 1.1.

Tabela1.1:Relações entre os modelos de riscos

Modelo	Restrição	Risco	Sobrevivência
PH/AFT	$e^{\alpha\prime_1 x} = \alpha_1, e^{\alpha_2' x} = \alpha_2$	$\alpha_1 h_0\left(\alpha_2 t\right)$	$\left[S_0\left(\alpha_2 t\right)\right]^{\frac{\alpha_1}{\alpha_2}}$
AFT	$e^{\alpha\prime_1 x} = e^{\alpha_2' x} = \alpha$	$\alpha h_0\left(\alpha t\right)$	$\left[S_0\left(\alpha t\right)\right]$
PH	$e^{\alpha\prime_1 x} = \alpha, e^{\alpha_2' x} = 1$	$\alpha h_0\left(t\right)$	$\left[S_0\left(t\right)\right]^{\alpha}$

1.3 Conteúdo da Dissertação

Uma introdução à análise de sobrevivência, alguns conceitos fundamentais e a importância do uso da função de risco para analisar dados de tempo de vida foram registrados anteriormente. Levando em consideração tal relevância, o objetivo da dissertação é apresentar um modelo de riscos proporcionais que leve em consideração o efeito do tempo no ajuste do modelo, mostrando propriedades e motivação para o uso do mesmo.

Nesta dissertação apresentamos, no Capítulo 2, os procedimentos de estimação pontual e intervalar em abordagem clássica. Esta análise clássica é baseada em propriedades assintóticas dos estimadores dos parâmetros envolvidos. Sendo assim, também foram realizados alguns estudos numéricos das estimativas clássicas e calculadas as probabilidades de cobertura dos intervalos de confiança, que também são apresentados neste Capítulo.

No Capítulo 3 apresentamos um estudo da metodologia de estimação intervalar e a construção de testes de hipóteses para os parâmetros do modelo de riscos proporcionais dependentes do tempo via reamostragem Bootstrap. Um estudo do custo ao estimar β na presença de α é apresentado no Capítulo 4. Uma análise Bayesiana para o modelo aqui proposto é apresentada no Capítulo 5.

Capítulo 2

Modelo de Risco Proporcional Dependente do Tempo

2.1 Introdução

Passa-se a apresentação de um modelo de risco proporcional dependente do tempo, baseado no modelo de risco proporcional de Cox, porém, caracterizado por aplicar um parâmetro que mede o efeito do tempo em sua própria estrutura.

A formulação do modelo e suas principais propriedades são apresentadas nas Seções 2.2 e 2.3, respectivamente.

Um procedimento de estimação pontual via máxima verossimilhança é mostrado na Seção 2.4, e, na Seção 2.5, são apresentados os testes de hipóteses assintóticos e o procedimento de estimação intervalar.

Ilustrando a metodologia apresentada neste capítulo, uma aplicação a dados reais é desenvolvida na Seção 2.6, e alguns estudos numéricos são descritos na Seção 2.7. As conclusões finais são apresentadas na Seção 2.8.

2.2 Formulação do Modelo

Denotando por T uma variável aleatória não negativa representativa do tempo de falha, temos que a função de risco é dada por,

$$h\left(t|\alpha,\beta\right)=\exp\left(\alpha t+\mathbf{x}^{'}\boldsymbol{\beta}\right), \tag{2.1}$$

em que α é uma medida de efeito do tempo e $\boldsymbol{\beta}^{'}=\left(\beta_1,.....,\beta_p\right)^{'}$ é um vetor de p parâmetros desconhecidos medindo a influência das p covariáveis $\mathbf{x}^{'}=\left(x_1,...,x_p\right)^{'}$. Neste modelo $e^{\alpha t}$ é uma função que descreve o risco de um indivíduo com $\mathbf{x}=0$ e $e^{\mathbf{x}^{'}\boldsymbol{\beta}}$ é o risco relativo, onde tem-se um aumento ou redução proporcional no risco associado com o conjunto de covariáveis. Observe-se que o aumento ou redução no risco é o mesmo em todo o tempo t.

Retornando a (2.1), e integrando de 0 a t obtem-se a função de risco acumulada, que é dada por,

$$H\left(t|\alpha,\boldsymbol{\beta}\right)=\frac{1}{\alpha}e^{\mathbf{x}^{'}\boldsymbol{\beta}}\left[e^{\alpha t}-1\right]. \tag{2.2}$$

De acordo com (2.1) a função de sobrevivência para variável aleatória T é dada por,

$$S\left(t|\alpha,\boldsymbol{\beta}\right)=\exp\left\{-\frac{1}{\alpha}\exp\left(\mathbf{x}^{'}\boldsymbol{\beta}\right)\left[\exp\left(\alpha t\right)-1\right]\right\}. \tag{2.3}$$

E ainda de (2.1) e (2.3), a função densidade de probabilidade de T é dada pela expressão:

$$f\left(t|\alpha,\beta\right)=\exp\left\{\alpha t+\mathbf{x}^{'}\beta-\frac{1}{\alpha}\exp\left(\alpha t+\mathbf{x}^{'}\beta\right)+\frac{1}{\alpha}\exp\left(\mathbf{x}^{'}\beta\right)\right\}. \tag{2.4}$$

Note-se que para $T>0$, tem-se que $S\left(0|\alpha,\beta\right)=1$ e correspondentemente $S\left(\infty|\alpha,\beta\right)=0$.

Vários modelos de risco usuais podem ser obtidos como casos particulares de(2.1) Quando $\alpha=0$ temos o modelo de riscos proporcionais de Cox dado em (Apêndice A) quando a função de risco base $h_0\left(t\right)=1$ (função de risco constante). Quando $\eta=e^{\mathbf{x}^{'}\beta}$ constata-se que o modelo segue a distribuição de Gompertz com parâmetros α e η (Apêndice B).

2.3 Propriedades do Modelo

A Figura 2.1 mostra a forma típica da densidade, função de risco e sobrevivência para um conjunto de parâmetros específicos ($\alpha = -0.05$ e $\beta = -1$).

A Figura 2.2 mostra curvas de risco típicas ilustrando a amplitude das formas que podem ser representadas, ao considerar-se valores específicos para os parâmetros do modelo. Observa-se que, quando α é positivo, a taxa de risco aumenta ao longo do tempo, enquanto que valores negativos diminuem o risco ao longo do tempo.

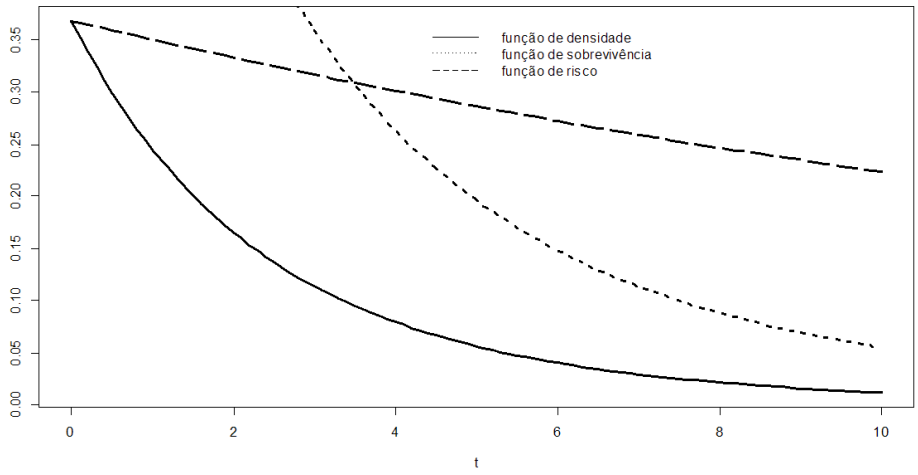

Figura 2.1: Funções de densidade, risco e sobrevivência

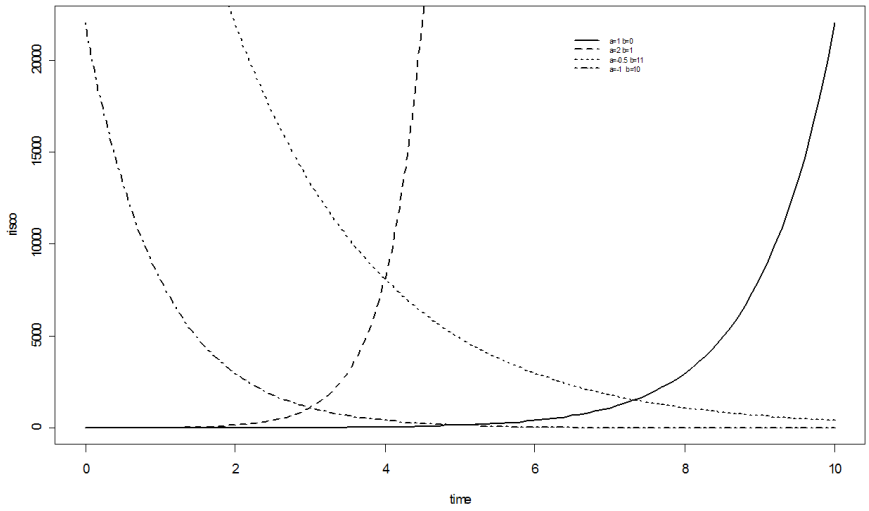

Figura 2.2: Várias formas para a função de risco

Considere-se o problema de duas amostras onde temos uma variável dummy $\mathbf{x}$ que serve para identificar Grupos 1 ou zero. Então tem-se que a função de risco é dada por,

$$h\left(t|\alpha,\beta\right)=\begin{cases} e^{\alpha t} & \text{se} \quad \mathbf{x}=0 \\ e^{\alpha t+\beta} & \text{se} \quad \mathbf{x}=1 \end{cases} \tag{2.5}$$

Assim, $e^{\alpha t}$ representa o risco no tempo t, grupo zero e $e^{\boldsymbol{\beta}}$ representa a razão do risco no grupo 1, relativo ao grupo zero para qualquer tempo t. Se $\boldsymbol{\beta}$ =0, os riscos são os mesmos nos dois grupos. Se β=0, 7, então o risco para um indivíduo no grupo 1, em qualquer tempo, é 2 vezes o risco de um indivíduo do grupo zero que tenha a mesma idade.

Na Figura 2.3, apresentam-se um exemplo construído para os indivíduos no Grupo 1, $\mathbf{x}=0$, e Grupo 2, $\mathbf{x}=1$, para o modelo dado pela equação (2.1), considerando $\alpha=1$ e $\beta=1$.

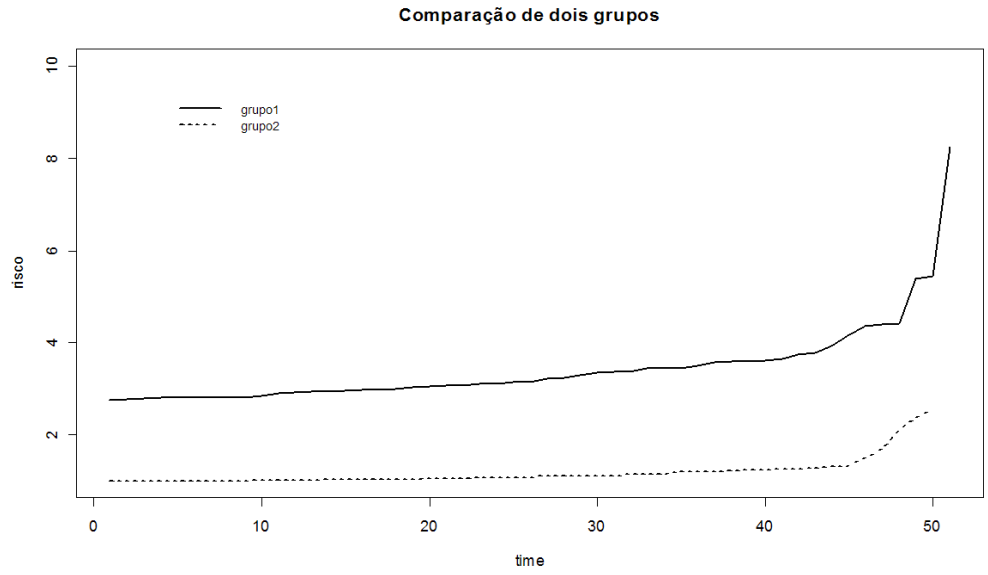

Figura 2.3: função de risco para os indivíduos no Grupo I e II

Nota-se que o modelo separa claramente o efeito do tempo do efeito das covariáveis, tomando o logaritmo, encontramos que o modelo (2.1) é um modelo aditivo com parâmetro dependente do tempo para o log do risco, isto é,

$$\log h\left(t|\alpha,\beta\right) = \alpha t + \mathbf{x}^{'}\boldsymbol{\beta}. \tag{2.6}$$

Como em todo modelo aditivo, assume-se que o efeito da covariável $\mathbf{x}$ é o mesmo em todo o tempo.

2.4 Procedimento de Estimação Pontual

Considere-se uma amostra aleatória de n indivíduos com as informações disponíveis (t_i, x_i, δ_i), em que $\delta_i = 1$, se o evento de interesse é observado e $\delta_i = 0$, caso contrário, para $i = 1, ..., n$. Então, a função de verossimilhança é dada por,

$$L\left(\alpha,\beta\right) = \prod_{i=1}^{n}\left[\exp\left(\alpha t_i + \boldsymbol{\beta}\mathbf{x}_i\right)\right]^{\delta_i}\left\{\exp\left[\left(-\frac{1}{\alpha}\right)\left(\exp\left(\alpha t_i + \boldsymbol{\beta}\mathbf{x}_i\right) - \exp\left(\boldsymbol{\beta}\mathbf{x}_i\right)\right)\right]\right\}. \tag{2.7}$$

Calculando o logaritmo da função de verossimilhança (2.7) temos,

$$l(\alpha,\beta)=\sum_{i=1}^{n}\delta_i\alpha t_i+\sum_{i=1}^{n}\delta_i\boldsymbol{\beta}\mathbf{x}_i-\sum_{i=1}^{n}\left(\frac{1}{\alpha}\right)\exp\left(\alpha t_i+\boldsymbol{\beta}\mathbf{x}_i\right)+\sum_{i=1}^{n}\left(\frac{1}{\alpha}\right)\exp\left(\boldsymbol{\beta}\mathbf{x}_i\right). \quad (2.8)$$

Os EMV são obtidos por derivação direta de (2.8). Os Estimadores de Máxima Verossimilhança (EMV) são obtidos por derivação direta de (2.8).

Derivando a equação (2.8) em relação à α temos,

$$\frac{\partial l}{\partial\alpha}=\sum_{i=1}^{n}\delta_i t_i+\frac{1}{\alpha}\left(\sum_{i=1}^{n}t_i\exp\left(\alpha t_i+\boldsymbol{\beta}\mathbf{x}_i\right)\right)+\frac{1}{\alpha^2}\left[\sum_{i=1}^{n}\exp\left(\alpha t_i+\boldsymbol{\beta}\mathbf{x}_i\right)-\sum_{i=1}^{n}\exp\left(\boldsymbol{\beta}\mathbf{x}_i\right)\right] \quad (2.9)$$

e derivando a equação (2.8) em relação à β_r , $r=1,...,k$ temos,

$$\frac{\partial l}{\partial\beta_r}=\sum_{i=1}^{n}\delta_i x_{ir}+\frac{1}{\alpha}\left[\sum_{i=1}^{n}\mathbf{x}_{ir}\exp\left(\boldsymbol{\beta}\mathbf{x}_{ir}\right)-\sum_{i=1}^{n}\mathbf{x}_{ir}\exp\left(\alpha t_i+\boldsymbol{\beta}\mathbf{x}_{ir}\right)\right], \quad (2.10)$$

Para resolver o sistema de derivadas $\frac{\partial l}{\partial\alpha}=0$ e $\frac{\partial l}{\partial\beta_r}=0$, utilizamos um método iterativo de Quasi-Newton (ver Nash, 1990).

Além disso, a matriz de informação observada I pode ser obtida através das derivadas segundas do log da função de verossimilhança. Os elementos da matriz de informação de Fisher são dados por (Lawless, 1982),

$$I_{jk}=E\left(-\frac{\partial^2 l}{\partial\beta_j\partial\beta_k}\right)=E\left(\frac{\partial l}{\partial\beta_j}\frac{\partial l}{\partial\beta_k}\right)=E\left(U_jU_k\right), \quad (2.11)$$

para $j,k=1,2$.

De acordo com (2.11) a matriz de informação de Fisher é, então, dada por,

$$I\left(\alpha,\beta\right)=\begin{pmatrix} E\left(U_\alpha U_\alpha\right) & E\left(U_\alpha U_{\beta_k}\right) \\ E\left(U_{\beta_k}U_\alpha\right) & E\left(U_{\beta_k}U_{\beta_k}\right)\end{pmatrix}. \quad (2.12)$$

2.5 Procedimento de Estimação Intervalar

Estimativas intervalares e testes de hipóteses para os parâmetros são baseados na distribuição normal assintótica dos estimadores de máxima verossimilhança (EMV) e na distribuição qui-quadrado da estatística da razão de verossimilhança (ERV), respectivamente (Lawless,1982). A adequação desses procedimentos é estudada na seção seguinte.

Seja $\hat{\theta} = (\hat{\alpha}, \hat{\beta})$ denotando os estimadores de $\theta = (\alpha, \beta)$. Assintoticamente, quando $n \to \infty$,

$$\hat{\theta} \to N(\theta, \mathbf{I}^{-1}(\theta)), \tag{2.13}$$

onde $\mathbf{I}$ é a matriz de informação esperada.

A distribuição assintótica dada em (2.13) pode ser utilizada para construir regiões de confiança aproximadas para os parâmetros e funções de interesse.

Para testar a variedade de hipóteses relacionadas com os parâmetros do modelo, podemos usar a estatística da razão de verossimilhança.

Sejam θ_0 e θ_1,denotando os EMV de θ sob as hipóteses H_0 e H_1. Sob certas condições de regularidade, a ERV é dada por,

$$\Lambda = -2 \log \left(\frac{L(\theta_0)}{L(\theta_1)} \right), \tag{2.14}$$

onde Λ tem uma distribuição assintótica qui-quadrado com d graus de liberdade (χ^2_d), em que d é a diferença entre o número de parâmetros independentes necessários para especificar H_0 e H_1.

2.6 Aplicação

Nesta Seção, serão apresentadas primeiramente, uma análise de dados completos segundo o modelo de risco proporcional dependente do tempo, e, em seguida, uma análise de dados censurados.

Exemplo1. Considerem-se os tempos de sobrevivência (em semanas) e contagem do número de glóbulos brancos de dois grupos de pacientes, Ag positivo e negativo, com diagnóstico de leucemia (Louzada-Neto *et al.*, 2002, pg 22).

Tabela 2.1: Dados de pacientes com leucemia

Tempos	Ag+	$\log_{10}$ (wbc)	Tempos	Ag-	$\log_{10}$ (wbc)
63	2300	3,36	56	4400	3,64
156	750	3,88	65	3000	3,48
134	2600	3,41	7	1500	3,18
16	6000	3,78	16	9000	3,95
108	10000	4,02	22	5300	3,72
121	10000	4,00	3	10000	4,00
4	17000	4,23	4	19000	4,28
39	5400	3,73	2	27000	4,43
143	7000	3,85	3	28000	4,45
56	9400	3,97	8	31000	4,49
26	32000	4,51	4	26000	4,41
22	35000	4,54	3	21000	4,32
1	100000	5,00	30	79000	4,9
1	100000	5,00	4	10000	5,00
5	52000	4,72	43	10000	5,00
65	100000	5,00			

Para todos os pacientes, a covariável contagem de glóbulos brancos(WBC) foi registrada na data do diagnóstico, estando seus respectivos logaritmos na base 10, apresentados também na referida tabela. As covariáveis presentes neste conjunto de dados são, portanto, X_1 = logaritmo da contagem de glóbulos brancos e X_2 = grupos (Ag+ ou Ag-).

Pode-se reescrever a equação (2.1) , considerando os dados da Tabela 2.1, da seguinte forma,

$$h\left(t|\alpha,\beta\right) = \exp\left(\alpha t + \beta_1 ag + \boldsymbol{\beta_2}\left(\log_{10}\left(wbc\right)\right)\right). \tag{2.15}$$

A Tabela 2.2 mostra as EMV, erro padrão E.P. e I.C. (95%) para o modelo (2.1). E a Figura 2.4 apresenta a função de risco, considerandos os grupos Ag+ e Ag-. Observa-se que o efeito das covariáveis é significativo, não acontecendo o mesmo em relação ao efeito

do tempo, devido ao fato de I.C. (95%) conter o valor 0.

Tabela 2.2: EMV, Erro Padrão (E.P.) e IC para dados Tabela 2.1

Parâmetros	EMV	E.P.	I.C.(95%)
α	-0.007	0.005	$[-0.018\ ;\ 0019]$
β_1	1.021	0.365	$[0.299\ ;\ 1.742]$
β_2	-0.968	0.087	$[-1.151\ ;\ -0.801]$

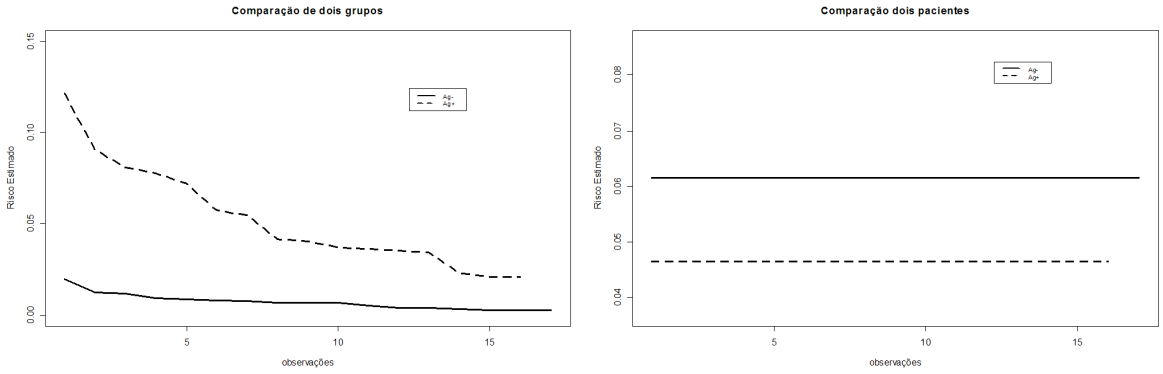

Figura 2.4: Comparação entre dois grupos (lado esquerdo) e dois pacientes (lado direito) para dados Tabela 2.1

Exemplo2. Com o intuito de incentivar a amamentação, pesquisadores do Departamento de Pediatria da UFMG fizeram um inquérito com mães de crianças com menos de 2 anos e que utilizavam o Centro de Saúde São Marcos localizado em Belo Horizonte. O estudo teve como objetivos principais conhecer a prática do aleitamento materno dessas mães bem como identificar possíveis fatores de risco ou de proteção para o desmame precoce. (Colosimo, E. A.; Giolo, S. R.; 2006)

A variável de interesse foi o tempo máximo de aleitamento materno (tempo contado a partir do mecanismo até o desmame completo da criança).

O conjunto de dados original é composto de 11 covariáveis mas após investigações preliminares, passa-se aqui a trabalhar com apenas 4 delas. Então, para todos os pacientes, as covariáveis presentes nesse conjunto de dados são, portanto, X1: experiência anterior de amamentação (0 se sim e 1 se não), X2: conceito materno sobre o tempo ideal de amamentação (0 se X2 >6 meses e 1 se X2 $\leq$ 6 meses), X3: dificuldades de amamentação

nos primeiros dias pós-parto (0 se nao e 1 se sim) e X4: recebimento exclusivo de leite materno na maternidade (0 se sim e 1 se não).

Pode-se reescrever a equação (2.1) considerando os dados da seguinte forma

$$h\left(t|\alpha,\beta\right) = \exp(\alpha t + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4). \tag{2.16}$$

Na Figura 2.5 , os gráficos mostram que a suposição de proporcionalidade não é violada.

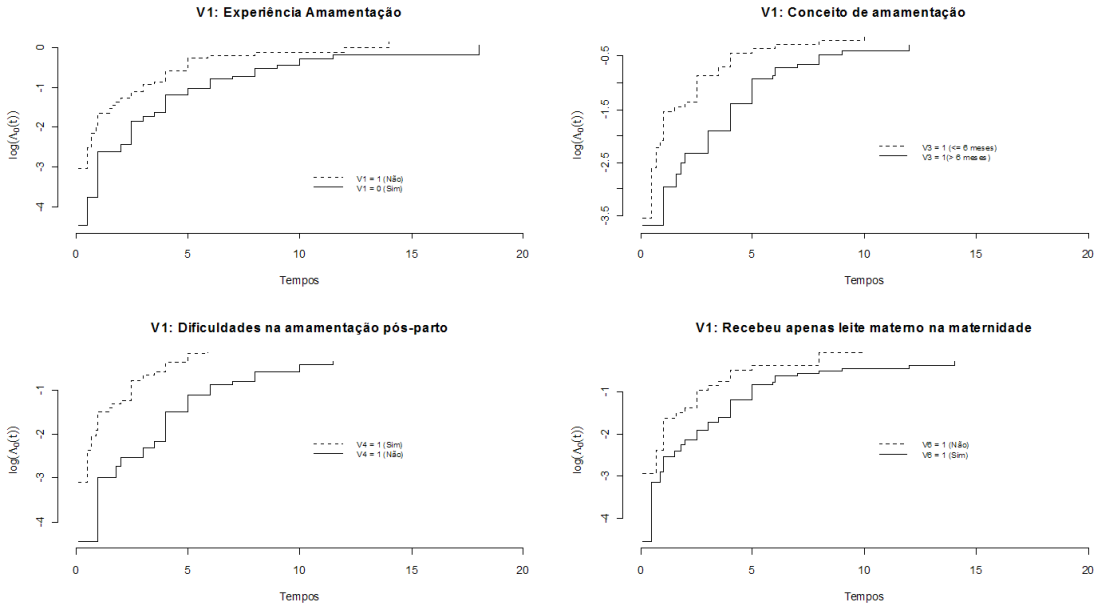

Figura 2.5.: Gráfico proporcionalidade

Utilizando o modelo de risco dependente do tempo (2.1) para a análise desses dados foram obtidos os resultados que estão condensados na Tabela 2.3.

Tabela 2.3: EMV, E.P.. e I.C. de 95% para equação (2.1)

Parâmetros	EMV	E.P.	I.C. (95%)
α	-0.3629	0.03594	$[-0.437; -0.296]$
β_1	-0.6128	0.1930	$[-1.004; -0.246]$
β_2	-0.7075	0.1876	$[-1.091; -0.354]$
β_3	-0.2879	0.1977	$[-0.689; 0.087]$
β_4	-0.4542	0.2067	$[-0.879; -0.066]$

O parâmetro que mede o efeito do tempo é significativo e apenas a covariável X3: dificuldades de amamentação nos primeiros dias pós-parto mostrou-se não significativa. O modelo pode ser expresso por:

$$h\left(t|\alpha,\beta\right) = \exp(-0.3629t + -0.6128x_1 + -0.7075x_2 + -0.2879x_3 + -0.4542x_4). \quad (2.17)$$

2.7 Estudo Numérico

São comuns, na prática, estudos com a possibilidade de amostras pequenas ou moderadas. Para avaliar a aplicabilidade dos resultados assintóticos nesses casos, as propriedades dos estimadores devem ser estudadas através do estudo da matriz de informação de Fisher.

Dentro das especificações gerais do estudo de simulação, os tempos de vida foram gerados segundo o método da inversa da função de distribuição (magalhães, 2004), onde, para os tempos,é assumida uma distribuição exponencial com taxa de falha 0.4 e 0.5, e as covariáveis foram geradas segundo uma Bernoulli com probabilidade de sucesso igual a 0.5. E então, foram gerados os tempos de falha, segundo o modelo com α = -0.5 e β = 1 para amostras de tamanhos n=10, 20, 30, 50 e 100 elementos.

Nesta seção, são apresentados os resultados de um estudo de simulação que verifica a probabilidade de cobertura dos intervalos de confiança assintóticos, tendo como parâmetro de variação o tamanho das amostras consideradas. A verificação do decaimento da variância dos EMV foi feita por meio da verificação do comportamento do $\log(var(.))$ versus $\log(n)$. Em uma regressão de $\log(var(.))$ versus $\log(n)$ o coefiente de inclinação deve ser aproximadamente -1. Os resultados do estudo são apresentados na Tabela 2.5, que mostra o decaimento. As probabilidades de cobertura dos intervalos de confiança de 95% são apresentadas na Tabela 2.4.

2.7.1 Cálculo da Probabilidade de Cobertura

Para a determinação da probabilidade de cobertura dos intervalos de confiança assintóticos gera-se uma amostra, definida como a amostra original, conforme as especificações já comentadas. E então, são obtidas as estimativas pontuais e intervalares. Em seguida,

foram geradas 1000 amostras a partir das mesmas distribuições utilizadas para a geração da amostra original, com os parâmetros destas distribuições substituídos pelos EMV, e verifica-se se as estimativas da amostra original estavam contidas ou não nos intervalos de confiança dessas 1000 amostras geradas. O número de vezes em que cada uma dessas situações ocorreu foi anotado. Assim, baseando-se nas 1000 amostras, calcula-se a probabilidade de cobertura dos intervalos de confiança assintóticos.

A Tabela 2.4 apresenta as probabilidades de cobertura dos intervalos de confiança de 95%.

Tabela 2.4. Probabilidade de Cobertura dos IC de 95%

n	α	β
10	0.931	0.979
30	0.991	0.996
50	0.948	0.908
70	0.965	0.988
100	0.986	0.965
300	0.988	0.986

Pela Tabela 2.4, concluí-se que a probabilidade de cobertura do intervalo de confiança de 95% tanto para o parâmetro α, quanto para o parâmetro β, é próxima da probabilidade de cobertura nominal de 95%, pelo menos para amostras moderadas.

Tabela 2.5. Coeficiente de Inclinação da relação entre o log da variância e n

n	$\log\left(var\left(\hat{\alpha}\right)\right)$	$\log\left(var\left(\hat{\beta}\right)\right)$
$10-50$	-1.614	-1.517
$50-70$	-0.667	-1.124
$70-100$	-0.5358	-0.6875
$100-300$	-1.153	-1.062

Os resultados da Tabela 2.5 mostram que as variâncias de α e β convergem para um com o aumento do tamanho amostral. Entretanto para amostras pequenas isto não acontece.

2.8 Considerações Finais

Uma das vantagens do modelo de riscos proporcionais dependente do tempo é permitir que o tempo seja considerado no ajuste do modelo. Vimos que o procedimento de estimação de máxima verossimilhança pode ser facilmente considerado, e que a estimação intervalar baseada na teoria assintótica pode ser considerada para amostras moderadas. Para amostras pequenas outro procedimento deve ser considerado.

Um método, já citado, muito utilizado atualmente, devido à facilidade computacional, para verificar resultados obtidos via teoria assintótica é o de reamostragem bootstrap que será tratado no Capítulo seguinte.

Capítulo 3

Método de Simulação *Bootstrap*

O método de reamostragem *bootstrap* foi desenvolvido por Efron (1979), sendo utilizado na obtenção de estimativas intervalares e pontuais, bem como na acurácia de testes e estimativas. No caso de amostras pequenas, o que ocorre com frequência em análise de sobrevivência, o *bootstrap* pode ser visto como uma técnica alternativa ao cálculo de intervalos de confiança.

A técnica de reamostragem bootstrap permite com precisão usar-se uma amostra para estimar a quantidade de interesse através de uma estatística e avaliar também as propriedades da distribuição dessa estatística, ou seja, fornece também estimativas para a distribuição, enviesamento, desvio padrão e intervalos de confiança da estatística (Canty, 2000). Como nos dias de hoje podemos contar com recursos computacionais intensivos não é necessário usar-se a teoria assintótica para análise das propriedades de uma estatística.

O processo de simulação de amostras *bootstrap* pode ser paramétrico ou não-paramétrico. No *bootstrap* paramétrico, utilizado quando se tem informação suficiente sobre a forma da distribuição dos dados, a amostra é formada realizando-se a amostragem diretamente nesta distribuição com os parâmetros desconhecidos substituídos por estimativas paramétricas. Já no *bootstrap* não-paramétrico, usado quando não se tem conhecimento da distribuição dos dados, são geradas R amostras com reposição e de mesmo tamanho da amostra original, à partir da distribuição $\hat{F}$, que corresponde a distribuição empírica dos dados.

Consedere $X_1, ..., X_n$ uma amostra aleatória de tamanho n com distribuição de probabilidade desconhecida F que depende de um parâmetro μ, e sejam $X = (x_1, ..., x_n)$ os valores observados. Como a função F é desconhecida, pode-se estimá-la pela função de

distribuição empírica, baseado na amostra n que é dada por,

$$\hat{F} = \frac{\#\left(x_i \leq x\right)}{n}.$$

A função $\hat{F}$ assume valores $1/n$ para cada valor da amostra x_i, $i = 1, ..., n$.

3.1 Intervalo de Confiança via Método *Bootstrap*

Em inferência estatística, tem-se interesse na quantificação do erro cometido ao se estimar um parâmetro de interesse θ através de $\hat{\theta}$. Uma estratégia usual para a busca de medidas de incerteza, que expressem esse erro, é a estimação do erro padrão de $\hat{\theta}$. Entretanto, métodos analíticos para a obtenção destas medidas nem sempre são disponíveis, ou constituem processos altamente complexos, enquanto métodos assintóticos, nos quais a construção de intervalo de confiança é baseada, dependem de aproximações nem sempre alcançadas. Neste contexto, o método *bootstrap* constitui uma eficiente alternativa, fornecendo estimativas do erro padrão de $\hat{\theta}$ livres de complexidades algébricas e possibilitando a obtenção de intervalos de confiança sem necessidade de pressupostos sobre a distribuição do estimador.

Desta forma, o método *bootstrap* é utilizado para a obtenção de estimativas intervalares empíricas para os estimadores dos parâmetros de interesse, através da reamostragem do conjunto de dados original.

Seja μ o parâmetro de interesse. Para cada amostra, calcula-se a EMV para μ, e tem-se no final de R reamostragens $\widehat{\mu}_1^* < ... < \widehat{\mu}_R^*$ valores das EMV ordenadas. Utiliza-se, então,

$$\widehat{\mu}^*_{(R+1)\left(\frac{\alpha}{2}\right)} \text{ e } \widehat{\mu}^*{}_{(R+1)\left(1-\frac{\alpha}{2}\right)}, \tag{3.1}$$

como sendo os limites inferiores e superiores do intervalo $100\,(1-\alpha)\,\%$ de confiança para μ. Em geral, o número de reamostragens R é fixado em 999, o que se considera aqui.

Desta forma, através de (3.1), podem ser obtidos intervalos de confiança percentil *bootstrap* $100\,(1-\alpha)\,\%$ para o parâmetro de interesse. Intervalos de confiança percentil *bootstrap* para os outros parâmetros de interesse são obtidos de maneira análoga.

Segundo Efron & Tibshirani (1993, p. 154), intervalos *bootstrap* também são aproxi-

mados, entretanto, oferecem melhores resultados que os intervalos de confiança padrão.

3.1.1 Aplicação

Considerando a metodologia acima foi desenvolvido um estudo de simulação *bootstrap,* adotados o modelo de risco dependente do tempo (1.1) e os dados referentes a aleitamento materno apresentados na seção 3.0.1. Esses dados foram reamostrados usando-se bootstrap não-paramétrico, em que os cálculos são repetidos 1000 vezes.

Na Tabela 2, encontram-se as EMV obtidas através da função *optim* do pacote R e os intervalos *bootstrap*. Observa-se, ainda, que os intervalos obtidos, considerado a técnica de reamostragem, estão próximos dos valores encontrados via método assintótico.

Tabela 3.1. EMV e I.C. de 95% *bootstrap*

Parâmetros	EMV	I.C. (95%)
α	-0.3137	$(-0.4671; -0.2844)$
β_1	-0.6536	$(-1.0027; -0.2999)$
β_2	-0.7207	$(-1.0929; -0.4161)$
β_3	-0.4155	$(-0.7016; 0.08078)$
β_4	-0.5669	$(-0.8552; -0.1511)$

3.1.2 Probabilidade de Cobertura dos Intervalos de Confiança *Bootstrap*

Para o cálculo da probabilidade de cobertura dos intervalos de confiança *bootstrap* gera-se uma amostra original utilizando o mesmo processo de geração da seção anterior. Através do procedimento *bootstrap* não-paramétrico geramos 399 réplicas da amostra original. As EMV de cada réplica foram calculadas. Cada uma das 399 réplicas foi replicada num número de 499 vezes. Assim, calcula-se os intervalos de confiança percentil *bootstrap.* Verifica-se o número de intervalos de confiança percentil *Bootatrap* que continham as EMV calculadas anteriormente. A probabilidade de Cobertura é dada pelo número de intervalos que contém as EMV dividido por 399.

A Tabela 5 apresenta as probabilidades de cobertura dos intervalos de confiança de 95%.

Tabela3.2. Probabilidade de Cobertura dos I.C. bootstrap de 95%

n	α	β
10	0.9373	0.9423
30	0.9649	0.9523
50	0.9598	0.9498
100	0.9699	0.9473
200	0.9774	0.9624
300	0.9824	0.9624

Desta forma, concluí-se que o intervalo de confiança *bootstrap* possue resultados próximos ao resultados obtidos via teoria assintótica (ver Tabela 2.4).

3.2 Considerações Finais

Utilizando-se o método de reamostragem Bootstrap, os intervalos de confiança obtidos via reamostragem se aproximam dos intervalos de confiança obtidos via teoria assintótica, provando assim a acurácia das estimativas intervalares calculadas via teoria assintótica. Entretanto, a utilização do procedimento Bootstrap é recomendado por independer da distribuição assintótica para ser construído.

Capítulo 4

O Custo de Estimar β na presença de α

4.1 Introdução

Em geral, para um modelo de regressão com n observações independentes, considera-se a variância da estimativa da quantidade de interesse sobre 2 cenários. Um deles é quando todos os parâmetros são estimados dos dados; e o outro é quando um subconjunto dos parâmetros é supostamente conhecido nos seus verdadeiros valores e os restantes são estimados.

Taylor, Siqueira e Weiss (1996) mostraram, sob certas condições, que a razão da soma acumulada através da projeção dos pontos da variância da quantidade de interesse é dada por q/p, onde q e p são o número de parâmetros livres nos dois cenários.

Assim, a inflação da variância associada com a adição de parâmetros, também no sentido do custo desta operação, é diretamente proporcional ao número de parâmetros.

Como provado anteriormente, o modelo de riscos proporcionais dependente do tempo pode ser visto como o modelo de Cox (1972), com um parâmetro a mais, que permite verificar o efeito do tempo na modelagem.

O que interessa, no momento, é medir o custo de estimar β, ao inserir-se o parâmetro α no modelo de riscos proporcionais , ou ainda, a influência quanto à variância da estimativa do efeito da covariável β é inflacionada por levar em consideração que α é estimado dos dados.

Primeiramente, na seção 4.2 foi apresentada uma medida para o custo, que é dada pela razão de variância, e, em seguida, na seção 4.3, um estudo numérico onde se fixa um valor para o parâmetro β e diversificando os valores do parâmetro α para vários tamanhos de amostras, a fim de obter-se valores da razão de variância, de acordo com tamanhos amostrais distintos.

4.2 Definindo o Custo de Estimar β

Para o cálculo da inflação da variância, Taylor, Siqueira e Weiss (1996) utilizaram um procedimento que é dividido em 2 estágios. O primeiro estágio consiste em estimar a quantidade de interesse, e o segundo em executar a análise condicionada ao valor estimado.

Os problemas em potencial na inferência associada ao estudo do custo já foram estudados por Miller (1990). A magnitude do problema vem sendo examinada para vários e diferentes modelos Taylor (1985), Faraway (1992) e Weiss (1995). E algumas sugestões envolvendo o método Bootstrap têm sido levantadas e consideradas (Faraway, (1992) e Efron and Gong, (1983)).

Bickel & Doksum (1981) estudaram o custo de estimar o parâmetro λ da transformação Box-Cox, definido como inflação da variância causada pela adição dos parâmetros. Carroll & Ruppert (1981) analisaram o custo de estimar a transformação do parâmetro λ para a mesma classe de modelos usados por Bickel & Doksum (1981), mas avaliados dentro de diferentes contextos. Taylor (1988) e Taylor *et al.*, (1996) estudaram o custo de adicionar parâmetros para modelos de resposta binária e Siqueira & Taylor (1999) avaliaram o custo de estimar λ da transformação Box-Cox na inferência do efeito de tratamento.

Neste trabalho, deseja-se saber qual o custo de estimar o efeito da covariável com a introdução de um parâmetro que mede o efeito do tempo, isto é, considerando o modelo dependente do tempo (2.1), determinar-se o custo de estimar β na presença de α.

Considere-se o vetor de parâmetros $\boldsymbol{\theta} = (\alpha, \beta)$ e sejam $\hat{\boldsymbol{\theta}} = \left(\hat{\alpha}, \hat{\beta}\right)$ as EMV. A matriz de informação esperada I pode ser obtida como em (2.12) . Suponha, ainda, que α é fixado pelo seu verdadeiro valor α_*; nesse caso a EMV é denotada por $\hat{\theta}_* = \hat{\beta}_*$. A dimensão da matriz de informação para esse caso é 1×1 e é denotada por I_*. Note que I_* é a matriz I dada em (2.12)sem a primeira linha e a primeira coluna.

A idéia principal é medir quanto à variância da estimativa do efeito da covariável ($\hat{\beta}$), inflacionada por levar em conta que α (o efeito do tempo) é estimada dos dados, isto é, o custo de estimar α na presença de β. Uma medida desse custo é dada pela razão de variância (RV), definida por,

$$RV(\overset{\wedge}{\beta}) = \frac{I_{22}^{-1}}{I_{*11}^{-1}}, \tag{4.1}$$

onde I_{ii}^{-1} é o i-ésimo elemento diagonal do inverso da matriz de informação esperada I para α e β avaliados em suas EMV $\hat{\alpha}$ e $\hat{\beta}$, e I_{*ii}^{-1} é o i-ésimo elemento diagonal do inverso da matriz de informação I_*.

Em geral, não há uma simples expressão analítica para $RV(\overset{\wedge}{\beta})$. Quando a razão variância $VR(\overset{\wedge}{\beta}) = 1$, conclui-se que não existe custo de estimar β na presença de α.

4.3 Estudo Numérico

Para visualizar o custo dado em (4.1) tempos de vida foram geradas segundo o método da inversa da função de distribuição onde para os tempos foi assumido uma distribuição exponencial com taxa de falha 0.4 e 0.5, respectivamente, as covariáveis segundo uma Bernoulli com probabilidade de sucesso igual a 0.5. Daí geramos os tempos de falha segundo o modelo com α = -10 e $\beta = 1$. Para o vetor de parâmetros $\boldsymbol{\theta} = (\alpha, \beta)$ assumimos $\alpha = -10, -5, -1, -0.5, -0.05, -0.01, 0.01,\ 0.05, 0.1$ e $\beta = 0.5$ também fixamos o número de eventos em $n = 10, 20, 50, 100$.

A Figura 1 apresenta as razões de variâncias obtidas neste estudo. Os resultados mostram que existe um baixo custo da estimativa de β, sob a influência de α, para valores negativos α. Quanto maiores são os α e mais se aproximam de zero a razão de variância tende a aumentar.

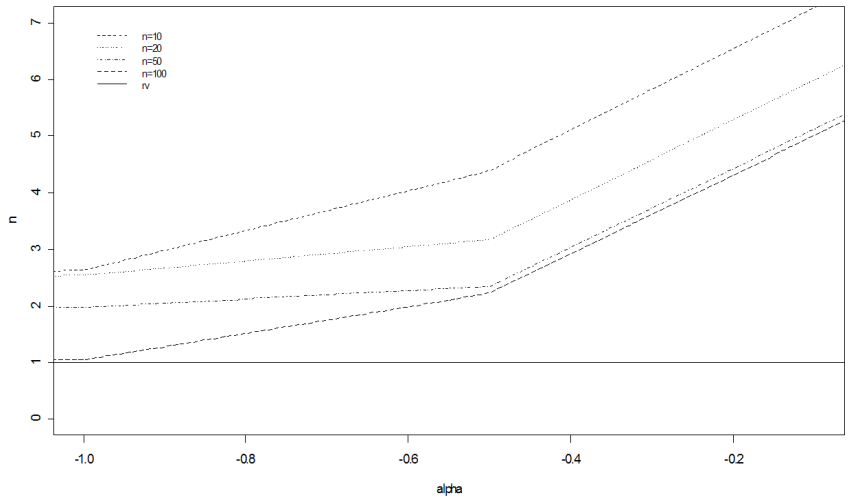

Figura 4.1- Gráfico da Razão de Variância

4.4 Conclusões

Neste Capítulo,foi obtida uma medida para a inflação da variância (RV) dada por (4.1), que também pode ser vista como uma medida do custo ao adicionar parâmetros a um modelo. Mediu-se o quanto a variância da estimativa do efeito da covariável β, no modelo de riscos proporcionais dependente do tempo, foi inflacionada, quando α é estimado dos dados.

Através de um estudo numérico mostrou-se que, para valores negativos de α, a RV é próxima de 1, pois o custo de estimar β na presença de α é baixo. Mas também vimos que, conforme α se aproxima de 0, a RV aumenta, ou seja, o custo de estimar β na presença de α aumenta, conforme α se aproxima de 0. E ainda que para valores maiores de n, como $n = 100$,a RV é mais próxima de 1. Conforme n diminui, a RV se distancia de 1, evidenciando um custo maior para tamanhos amostrais menores.

Sendo assim cé obvio que, para valores negativos de α, os resultados são convenientes. Mas, para valores positivos de α, e α próximo de zero, os resultados não são convenientes.

Capítulo 5

Abordagem Bayesiana

5.1 Introdução

Após o estudo para o modelo de riscos proporcionais dependente do tempo (2.1) , sob o ponto de vista frequentista,é importante considerarmos a abordagem Bayesiana, na qual é combinada a opnião de especialistas com a informação extraída dos dados. Tratar-se-á, pois, a seguir, sobre a metodologia Bayesiana, em todas as suas implicações, para o modelo em pauta.

5.2 Metodologia

O interesse básico de um estudo estatístico está em obter informações sobre uma quantidade de interesse θ. A análise Bayesiana de dados pode ser vista como um procedimento para fazer inferência dos dados, usando modelos probabilísticos para quantidades observadas e outras de que se deseja tomar conhecimento (ver Gelman, Carlin e Rubin, 2003). Neste contexto, é possível que pesquisadores atribuam modelos distintos para θ.

Teorema de Bayes

Seja θ uma quantidade de interesse desconhecida. A informação a *priori* a respeito desta quantidade é representada aqui por $p(\theta)$. E a informação que temos dos dados é representada pela função de verossimilhança $p(x|\theta)$. O teorema de Bayes traduz a idéia de que, depois de se observar $X = x$, a quantidade de informação sobre θ é aumentada (ver Eshler, 2003).

O teorema de Bayes é dado por,

$$p(\theta|x) = \frac{p(\theta, x)}{p(x)} = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{p(x|\theta)p(\theta)}{\int p(\theta, x)d(\theta)}. \tag{5.1}$$

Note-se que $\frac{1}{p(x)}$, não dependente de θ, funciona como uma constante normalizadora de $p(\theta|x)$.

A função de verossimilhança $p(x|\theta)$ fornece a *plausibilidade* de cada um dos possíveis valores de θ, e a função $p(\theta)$ é chamada de distribuição à *priori.*

A forma usual para o teorema aqui apresentado é

$$p(\theta|x) \propto p(x|\theta)p(\theta). \tag{5.2}$$

A combinação da informação a *priori* com a informação dos dados, traduzida pela verossimilhança, dá a informação à *posteriori* de θ que é convenientemente resumida em termos de esperanças de funções particulares do parâmetro θ,isto é,

$$E\left[g\left(\theta\right)|x\right] = \int g\left(\theta\right) p\left(\theta|x\right) d\theta. \tag{5.3}$$

Assim, o problema geral da inferência Bayesiana consiste em calcular tais valores esperados segundo a distribuição a *posteriori* de θ.

Para a resolução das questões mencionadas acima, métodos numéricos vêm sendo desenvolvidos nas últimas décadas. Os mais utilizados são: métodos de integração simples de Monte Carlo, os métodos de reamostragem por importância e os métodos de Monte Carlo via Cadeias de Markov (MCMC). Estes são mais simples para implementação e não apresentam restrições quanto ao número de parâmetros a serem estimados.

5.2.1 Cadeis de Markov Monte Carlo (MCMC)

O método MCMC é uma forma de integração Monte Carlo. A idéia é simular uma cadeia de Markov irredutível e aperiódica cuja distribuição estacionária é a distribuição de interesse $\pi\left(\theta\right)$. Para os bayesianos isso é a densidade a *posteriori* $p\left(\theta|x\right)$. Existem 2 métodos de gerar cadeia de Markov com distribuição estacionária especificada. Um deles que tem sido usado por muitos anos em estatística física é o agoritmo de metropolis (Metropo-

lis *et al.*, 1953). Hastings (1970) mostra uma generalização do algoritmo Metropolis. O outro método é o Gibbs Sampler (Geman & Geman,1984) que foi traduzido dentro da influência da literatura por Gelfand & Smith (1990). É importante ressaltar que o uso desses algoritmos, em geral, é necessário se a geração não-iterativa da distribuição da qual se deseja obter uma amostra for muito complicada ou custosa.

Na subseção será apresentado o algorítmo de Metropolis-Hastings, utilizado, neste trabalho para obter as estimativas dos parâmetros de interesse do modelo de riscos proporcionais dependente do tempo dado em (2.1) .

5.2.2 Metropolis-Hastings

Quando as distribuições condicionais a *posteriori* não são facilmente identificadas como possuidoras de uma forma padrão (normal. gama, etc.), o que impossibilita a geração direta a partir destas distribuições, usa-se o algoritmo Metropolis (Metropolis *et al.*, 1953). Uma generalização foi dados por Hastings (1970) e iremos descrevê-la aqui. A idéia dos algorítmos de Metrópolis-Hastings é, um valor é gerado de uma distribuição auxiliar e aceito com uma dada probabilidade. Esse mecanismo de correção garante que a convergência da cadeia para a distribuição de equilíbrio, que neste caso é a distribuição à posteriori (ver Ehlers, 2003).

Supondo que a cadeia esteja no estado θ e um valor θ' é gerado de uma distribuição proposta $t(.|\theta)$. Note-se que a distribuição proposta pode depender do estado atual da cadeia, por exemplo $t(.|\theta)$ poderia ser uma distribuição normal centrada em θ. O novo valor θ' é aceito com probabilidade

$$\alpha(\theta, \theta') = \min\left(1, \frac{\pi(\theta')\, t(\theta, \theta')}{\pi(\theta)\, t(\theta', \theta)}\right). \tag{5.4}$$

onde π é a distribuição de interesse.

Em termos práticos, o algoritmo de Metropolis-Hastings pode ser especificado pelos seguintes passos,

1 - Inicie o contador de iterações i = 0 e especifique um valor inicial $\theta^{(0)}$.

2 - Gere um novo valor θ' da distribuição $t(.|\theta)$.

3 - Calcule a probabilidade de aceitação $\alpha(\theta, \theta')$ e gere $u \tilde{} U(0, 1)$.

4 - Se $u \leq \alpha$ então aceite o novo valor e faça $\theta^{i+1} = \theta'$, caso contrário rejeite e faça $\theta^{i+1} = \theta$.

5 - Incremente o contador de i para i + 1 e volte ao passo 2.

5.2.3 Diagnósticos de Convergência

Os métodos MCMC constituem uma ótima ferramenta para a resolução de problemas práticos mas atenção especial deve ser dada para algumas dificuldades que podem vir a surgir. Podemos citar aqui: o número de iterações para se obter a convergência, a possibilidade das iterações iniciais da amostra serem influenciadas pelos valores iniciais dos parâmetros e ainda, das sequências de valores apresentarem correlação entre os parâmetros.

Os métodos de verificação de convergência são baseados nas propriedades da cadeia de Markov e indicam a convergência ou não da amostra simulada para a distribuição marginal.

A princípio, com a intenção de verificar a convergência, pode ser feita uma análise dos gráficos ou das medidas descritivas (média, desvio-padrão e os quantis) obtidas a partir dos valores simulados para os parâmetros de interesse. Pode-se destacar entre os gráficos mais frequentes, o da quantidade de interesse estimada ao longo das iterações e o da estimativa da distribuição marginal a *posteriori* deste parâmetro.

Outra avaliação da convergência é feita, utilizando-se algumas técnicas de diagnóstico das cadeias. As técnicas de diagnóstico mais populares são desritas por Geweke (1992), Gelman & Rubin (1992) e Raftery e Lewis (1992). Estas técnicas estão implementadas na biblioteca **CODA (Convergence Diagnosis and Output Analysis Software for Gibbs Sampling Output)** (Best et al., 1997).

Nenhum desses métodos é considerado mais eficaz que o outro então, são utilizados conjuntamente a fim de obter-se uma indicação de convergência.

Algumas técnicas de identificação e monitoração formal e informal de convergência foram revisadas por Gamerman (1991). E ainda, Brooks e Roberts (1995) revisaram alguns dos métodos de convergência e discutiram a implementação destes e suas possíveis extensões.

5.3 Aplicação

Desenvolvimento do procedimento Bayesiano para o modelo, considerando -se que as distribuições a priori para α e β são independentes e dadas por,

$$\pi(\alpha) = \frac{1}{\sqrt{2\pi}\sigma_\alpha} \exp\left\{-\frac{1}{2\sigma_\alpha^2}(\alpha - \mu_\alpha)^2\right\}, \tag{5.5}$$

e

$$\pi(\beta) = \frac{1}{\sqrt{2\pi}\sigma_\beta} \exp\left\{-\frac{1}{2\sigma_\beta^2}(\alpha - \mu_\beta)^2\right\}. \tag{5.6}$$

Distribuições a Posteriori Conjunta

$$\pi(\alpha, \beta|Dados) \propto \exp\left\{\begin{array}{c} \frac{-(\alpha-\mu_\alpha)^2}{2\sigma_\alpha^2} - \frac{(\beta-\mu_\beta)^2}{2\sigma_\beta^2} + [\sum_{i=1}^n \delta_i(\alpha t_i + \beta x_i)] - \\ -\frac{1}{\alpha}\sum_{i=1}^n [\exp(\alpha t_i + \beta x_i) - \exp(\beta x_i)] \end{array}\right\}. \tag{5.7}$$

Distribuições a Posteriori Condicionais

$$\pi(\alpha|\beta, Dados) \propto \exp\left\{\sum_{i=1}^n \delta_i(\alpha t_i) - \left(\frac{1}{\alpha}\right)\sum_{i=1}^n [\exp(\alpha t_i + \beta x_i) - \exp(\beta x_i)]\right\}\pi(\alpha) \tag{5.8}$$

$$\pi(\beta|\alpha, Dados) \propto \exp\left\{\sum_{i=1}^n \delta_i(\beta x_i) - \left(\frac{1}{\alpha}\right)\sum_{i=1}^n [\exp(\alpha t_i + \beta x_i) - \exp(\beta x_i)]\right\}\pi(\beta) \tag{5.9}$$

Valores dos hiperparâmetros, aleatórios, fixados para o exemplo em questão,

$$\pi(\alpha) = N(0, 1),$$

e

$$\pi(\beta) = N(0, 1).$$

As densidades a *posteriori* marginais da distribuição a *posteriori* não são facilmente

obtidas, isto porque a integração da densidade a *posteriori* conjunta é complicada. Opta-se aqui por trabalhar com o software **R** (Spiegelhalter *et al,* 1997) para a geração das distribuições a *posteriori* dos parâmetros do modelo dado por (2.1) e cálculo das estimativas de interesse.

Exemplo Dados Completos: Toma-se mão aqui de dados simulados utilizados anteriormente neste trabalho. Foram geradas 2 cadeias de 200.000 iterações para os parâmetros. As primeiras 100.000 foram deprezadas. Os resultados obtidos estão apresentados na Tabela 5.1.

Tabela 5.1 - Estatísticas Resumo das distribuições a posteriori no Modelo dado em (2.1), utilizando a distribuição a Priori $N(0,1)$

Parâmetro	Média	E.P.	I.C.(95%)
α	1.3353	0.1891	$(0.9508; 1.6975)$
β	-0.8095	0.23	$(-1.2715; -0.3715)$

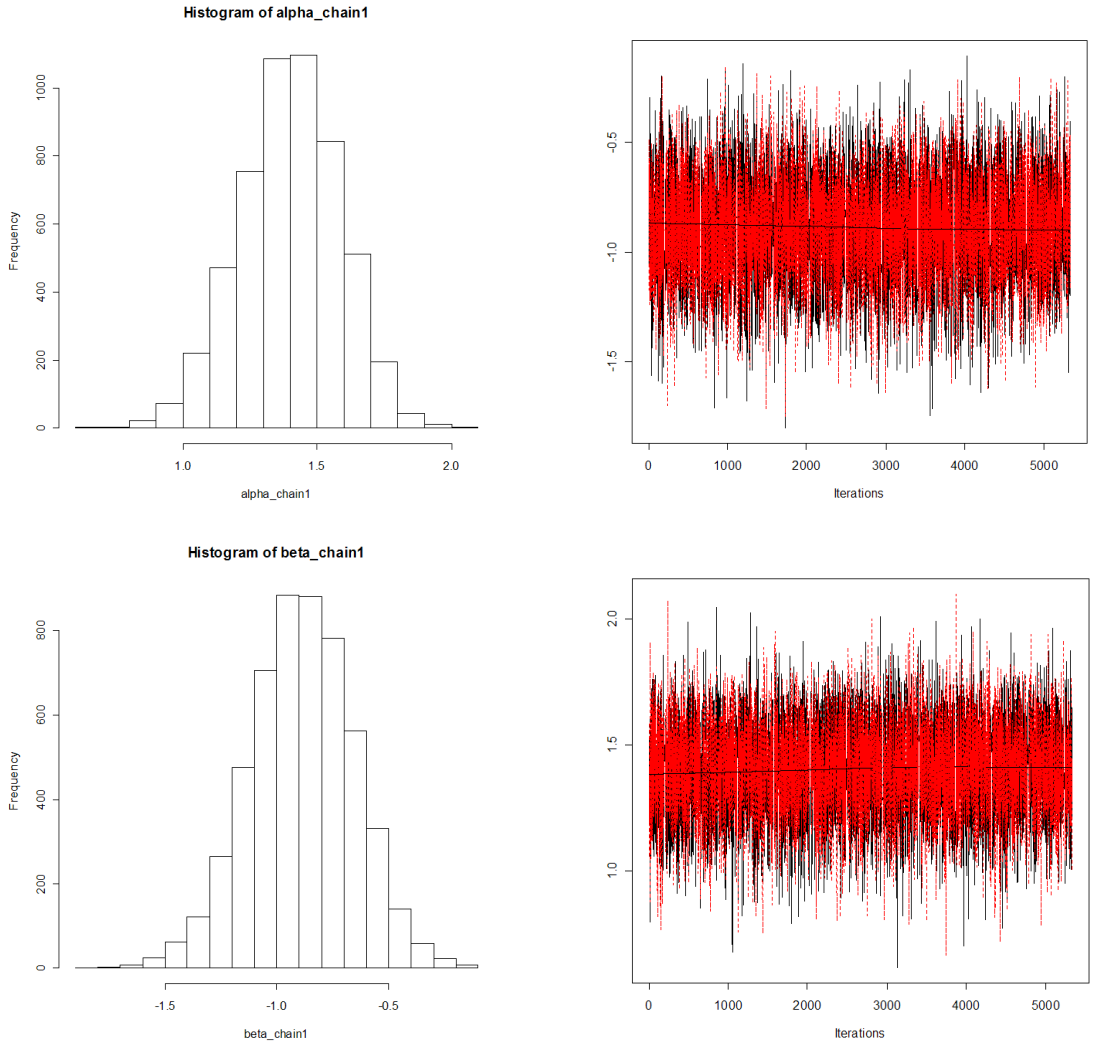

Figura 5.1. Posteriori Marginal e traço da cadeia para α (acima) e β (abaixo).

Pelos resultados apresentados na tabela 5.1 nota-se que para o conjunto de dados aqui utilizado é significativo o efeito do tempo e da covariável presente. A convergência das cadeias geradas está de acordo com os diagnósticos de convergência implementados no CODA (Best *et al.,* 1997). Os traços das cadeias e a estimação da densidade para cada parâmetro, apresentados na Figura 5.1, indicam que não há problemas com a convergência do algoritmo.

Exemplo Dados Censurados: Neste exemplo foram utilizados os dados apresentados anteriormente referentes a aleitamento materno (ver Colosimo, E. A.; Giolo, S. R.; 2006), considerando uma única covariável X_1 = experiência anterior de amamentação.

Tabela 5.2 - Estatísticas Resumo das distribuições a posteriori no Modelo dado em (2.1), utilizando a distribuição a Priori $N(0,1)$

Parâmetro	Média	E.P.	I.C.(95%)
α	-0.4593	0.0369	$(-0.5335; -0.3915)$
β	-0.9389	0.1679	$(-1.2724; -0.6255)$

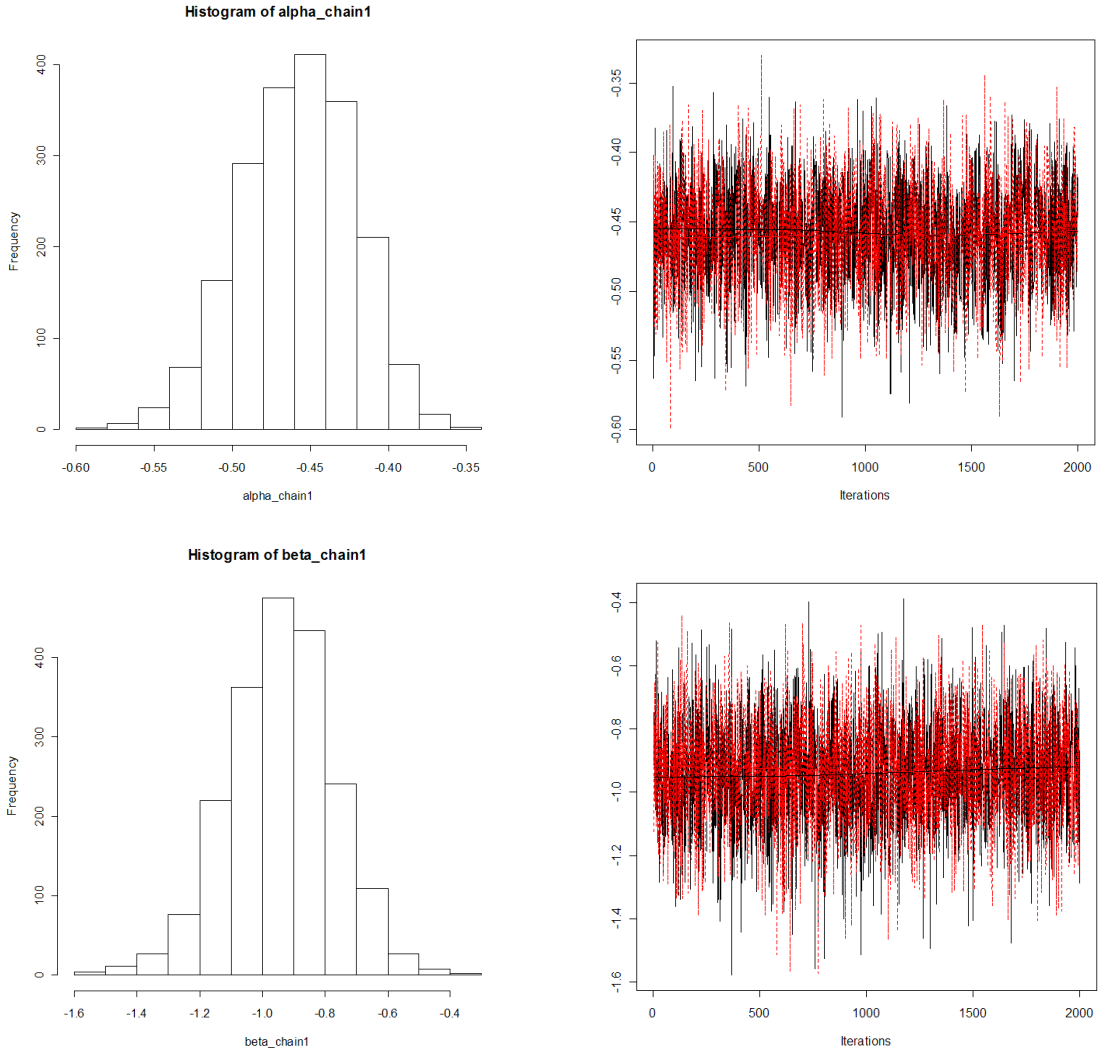

Figura 5.2. Posteriori Marginal e traço da cadeia para α (acima) e β (abaixo).

Pelos resultados apresentados na tabela 5.2 nota-se que para o conjunto de dados

referente a aleitamento materno é significativo o efeito do tempo e da covariável experiência anterior com leitamento materno presente no conjunto de dados. A convergência das cadeias geradas está de acordo com os diagnósticos de convergência implementados no CODA (Best *et al.,* 1997). Os traços das cadeias e a estimação da densidade para cada parâmetro, apresentados na Figura 5.2, indicam que não há problemas com a convergência do algoritmo.

Capítulo 6

Conclusões e Perspectivas Futuras

Em razão de minusciosa análise dos métodos assintóticos, considerações sobre reamostragem, estudo da inflação da variância, e reflexões do ponto de vista *Bayesiano,* propõe-se, finalmente, um modelo de risco proporcional dependente do tempo em sua modelagem.

Na abordagem clássica, verifica-se que não há problemas referentes a estimação via máxima verossimilhança e, para o conjunto de dados aqui estudado, mostra-se que o efeito do tempo é significativo e, ainda, a estimação de máxima verossimilhança pode ser considerada para tamanhos amostrais moderados. Mas deve ser registrado que é necessária uma certa cautela, ao se utilizar resultados assintóticos encontrados, pois a teoria clássica usual pode não ser válida, uma vez que, em análise de sobrevivência, é comum a aplicação de amostras pequenas. Foram utilizadas de técnicas de reamostragem na obtenção de estimativas pontuais e, ainda, da acurácia dos valores obtidos via teoria assintótica. Os intervalos de confiança obtidos via teoria assintótica se aproximam dos intervalos de confiança encontrados no método de simulação Bootstrap, garantindo assim, a eficiência do procedimento. Ressalta-se que o método *Bootstrap* é recomendado por independer de distribuição assíntótica para ser construído. Na análise do custo de estimar β na presença de α, mostra-se que, conforme α se aproxima de 0, a RV aumenta. E ainda que, para valores maiores de n, como $n = 100$,a RV é mais próxima de 1. Conforme n diminui, a RV se distancia de 1, evidenciando um custo maior para tamanhos amostrais menores.

Do ponto de vista Bayesiano observa-se que não existem dificuldades de implementação do método proposto e, também, são alcançados bons resultados nas aplicações feitas aqui,

mostrando assim outra metodologia apropriada para o estudo do modelo aqui proposto.

O modelo de riscos proporcionais dependente do tempo mostrou-se útil para resolver problemas que eventualmente podem ocorrer quando os dados de sobrevivência são proporcionais. Uma análise do modelo pode conduzir a vários estudos futuros, como por exemplo, considerar o modelo com um termo de fragilidade presente.

Referências Bibliográficas

[1] ANDERSEN, P.K. and GILL, R.D.(1982).Cox's Regression Model for Counting Process: a Large Sample Study. The Annals of Statistics,10,1100-1120.

[2] BEST, N., COWLES, M. AND VINES, K. (1997). CODA - Convergence Diagnosis and Output Analysis Software for Gibbs Sampling Output: Version 0.4, MRC Biostatistics Unit, Cambrigde UK.

[3] BICKEL, P.J., DOCKSUM, K.A.(1981).In Analysis of Transformations Revisited. Journal of the American Statistical Association,76,296-311.

[4] BROOKS, S. P.; ROBERTS, G. O (1998). Statistics and Computing. Statslab.com.ac.uk

[5] CARROL, R.J. and RUPERT, D.(1981).On Prediction and the Power Transformation Family. Biometrika,68,609-615.

[6] CHALITA, L. V. A. S., COLOSIMO, E. A., DEMÉTRIO, C. G. B., BARBIN, D. e SIMÃO, S.,(1999).Modelos de Regressão para dados de Sobrevivência Agrupados Aplicados a um Estudo Agronômico. Revista de Matemática e Estatística, 17, 193-207.

[7] Canty, A. J.; Tibshirani, R. J.; Leger, C. (2000). The estimation function Bootstrap, JSTOR.

[8] CIAMP, A. and J. ETAZADI-AMOLI. A general Model for testing the proportinal hazards and the acelerated failure time hypothesis in analysis of censored survival data with covariates. Commun.Statist. A, 14, 651-667 (1985).

[9] COLOSIMO, E. A.; GIOLO, S. R. (2006).Análise de Sobrevivência Aplicada. ABE - Projeto Fisher. Editora Blucher.

[10] COLLETT, D. (1994). Modelling Survival Data in Medical Research. Chapman and Hall, New York.

[11] COX, D.R.(1972a). Regression Models and Life-tables(with discussion). Journal of Royal Statistical Society,B,34,187-220.

[12] COX, D. R. and HINKLEY, D.V.(1974).Theorical Statistics.London:Chapman & Hall.

[13] COX, D. R. (1975). Partial Likelihood, Biometrika, 62.

[14] COX, D. R. and OAKES, D. (1984). Analysis of Survival Data. London: Chapman & Hall.

[15] CREMASCO, C. P.. Modelagem de Dados de Sobrevivência via Modelo de Risco Logístico Generalizado, Dissertação Apresentada ao Departamento de Estatística da Universidade Federal de São Carlos, Março de 2005.

[16] DAVISON, A.C.,HINKEY, D.V.(1997).Bootstrap Methods and their Application.Cambridge: Cambridge University Press,582p.

[17] EFRON, B. (1979). Bootstrap Methods: another look at the jacknife. Annals of Statistics, 7, 1-26.

[18] EFRON, B. and TIBHIRANI, R.J.(1993). An Introduction to the Bootstrap. New York: Wiley,436p.

[19] EHLERS, R. S. (2006). Introdução a Inferência Bayesiana. Departamento de Estatística - UFPR.

[20] GARG, L.M.; RAO, B.R.; REDMOND, C.K. (1970). Maximum-likelihood Estimationof the Parameters of the Gompertz Survival Function. Applied Statistics, vol. 19, 2, 152-159.

[21] GAMERMAN, D. (1991). Dynamic Bayesian Models for Survival Data. Applied Statistics 40, 63 - 79.

[22] GEMAN, S. and Geman, D. (1984). Stocastic Relation, Gibbs Distribution and the Bayesian restoration of images. IEEE Transction on Pattern Analysis and Machine Inteligence, 6, 721 - 741.

[23] GELFAND, A. E., Dey, D. K. and Chang, H. (1992). Model determinating using predictive distribution with implementation via sampling-based methods. (with discution). In Bayesian Statistics, 4.

[24] GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2003). Bayesian Data Analysis. London: Chapman & Hall.

[25] GEISER, P.W.; CHANG, M.N.; RAO, P.V.; SHUSTER, J.J; PULLEN, J. (1998). Modelling Cure Rates Using Gompertz Model With Covariate Information. Statist. Med. 17, 831 - 839.

[26] GELFAND, A. E., DEY, D. K. and CHANG, H. (1992). Model determinating using predective distribution with implementation via sampling-based methods. In Bayesian Statistics, 4.

[27] GEWEKE, J. (1989). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In Bayesian Statistics, 4.

[28] GOMPERTZ, B.(1825). On the Nature of the Function Expressive of the Law of Human Mortality and on a New Mode of Determining Life Contingencies. Philos, Trans.Roy.Soc. London A 115, 513-585.

[29] GREEN,M.S; e SYMONS,M.J.; A comparasion of the logistic risk function and the proportional hazards model in prospective epidemiologic studies,1983.Journal of Chronic Disease, Vol. 36, 10, 715-724.

[30] HALL, P. and WILSON,S.R.(1991).Two Guidelines for Bootstrap Hipothesis testing. Biometrics, 47, 757-762.

[31] HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their aplications. Biometrika, 57, 97 - 109.

[32] HIGGINS, T.; Mathematical Models of Mortality. Paper Presented at the Workshop on Mortality Model ling and Forecasting. Australian National University. Febuary, 2003.

[33] KALBFLEISH, J.F, PRENTICE, R.L. The Statistical Analisys of Failure Time Data. John Wiley and Sons, New York, 1980.

[34] LAWLESS,J.F. Statistical Models and Methods for Lifetime Data. New York: John Wiley & Sons, 1982.

[35] LEE, E.T. Statistical Methods For Survival Data Analysis. John Wiley and Sons, New York, 1992.

[36] LOUZADA-NETO, F. (1997). Extend Hazard Regression Model for Reability and survival analysis. Lifetime Data Analysis, 3, 367-381.

[37] LOUZADA NETO,F; PERDONÁ, G; Accelered Lifetime Tests with a Log-Non-Linear Stress-Response Relationship for Reability Data. Relatório Técnico do Departamento de Estatística da Universidade Federal de São Carlos. Agosto, 2000.

[38] LOUZADA NETO, F.; MAZUCHELI, J. ; ACHCAR, J. A; Introdução a análise de sobrevivência e Confiabilidade. IMCA,2002.

[39] MACKENZIE, G.; A Logistic Regression Model for Survival Data. In: 17th International Workshop in Statistical Modelling, p. 431-438.

[40] METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M., TELLER, A. H. and TELLER, E. (1953). Equations of State calculations by fast computing machine. Journal of Chemical Physics, 21, 1087 - 1091.

[41] NASH, J. C. (1990) Compact Numerical Methods for Computers. Linear Algebra and Function Minimisation. Adam Hilger.

[42] NELDER,J.A.,WEDDERBURN, W.M.,(1972).Generalized Linear Models. Journal of Royal Statistical Society, A, 135, Part 3, 370-384.

[43] PRENTICE, R.L.; SHAARAWI, E., (1973). A Model for Mortality Rates anda a Test of Fit for the Gompertz Force of Mortality.Applied Statistics, Vol. 22, No. 3 , pp. 301-314.

[44] RAFTERY, A. E., and LEWIS, S. (1992). How many iterations in the Gibbs Sampler? In Bayesian Statistics, 4.

[45] SIQUEIRA, A. L. and TAYLOR , J.M.G.(1999).Treatmen Effects in a Logistic Model Involving the Box-Cox Transformation. Journal of the American Statistical Association, 94, 240-246.

[46] TAYLOR, J.M.G.; SIQUEIRA, A. L. and WEISS, R.E.(1996). The Cost of Adding Parameters to a Models. Journal of Royal Statistical Society, B, 58, 593-680.

[47] TAYLOR, J.M.G (1988). The Cost of Generalizing Logistic Regression. Journal of the American Statistical Association, 87, 1078-1083.

[48] TOMAZELLA, V. L. D; Modelagem de Dados de Eventos Recorrentes via Processo de Poisson com Termo de Fragilidade. Tese Apresentada ao Instituto de Ciências Matemáticas e Computação - ICMC, Julho de 2003.

[49] WILLEKENS, F.(2001). Gompertz in context: The Gompertz and related distributions. Forecasting mortality in developed countries. Kewrler Academic Publishers, TheNetherlands, 105-126.

Apêndice A

O Modelo de Gompertz

Benjamim Gompertz em 1825 sugeriu uma lei de progressão geométrica para a mortalidade de um indivíduo depois de uma certa idade, ou ainda, uma lei de mortalidade capaz de explicar como o tempo influencia na morte de um indivíduo. Willekins (2001) mostrou que a função de distribuição de Gompertz é do tipo valor extremo.

A função de risco de Gompertz é dada por,

$$h\left(t\right)=\beta\exp\left(\alpha\mathbf{t}\right) \tag{A.1}$$

e ainda,

$$S\left(t\right)=\exp\left\{-\left(\frac{\beta}{\alpha}\right)\left(e^{\alpha t}-1\right)\right\}, \tag{A.2}$$

então por $(A.1)$ e $(A.2)$ temos que :

$$f\left(t\right)=\beta e^{\alpha t}\exp\left\{-\left(\frac{\beta}{\alpha}\right)\left(e^{\alpha t}-1\right)\right\}, t\geq 0. \tag{A.3}$$

Note que o $\log h\left(t\right)$ é uma função linear da idade

$$h\left(t\right)=\log(\beta)+\alpha\mathbf{t}. \tag{A.4}$$

A.1 A Função de Verossimilhança de Gompertz

Segundo Garg, Raja Rao e Redmond (1970), ao se levar em consideração um intervalo de tempo $(0, t_p)$ que pode ser dividido em p subintervalos da seguinte forma $(0, t_1)$, $(t_1, t_2), ..., (t_{p-1}, t_p)$,e sejam

n = número de indivíduos na amostra,

d_i = número de falhas observadas no intervalo de tempo (t_{i-1}, t_i),

s_i = número de censuras até o tempo t_i, $i = 1, ..., p$.

A probabilidade de d_i mortes no intervalo (t_{i-1}, t_i) é $\{S(t_{i-1}) - S(t_i)\}^{d_i}$ que pode ser aproximado por $\{f(\tau_i)\}^{d_i}$, onde τ_i é o ponto médio do intervalo (t_{i-1}, t_i).

Feitas estas considerações a função de verossimilhança para o modelo de Gompertz pode ser escrita da seguinte forma,

$$L(\mathbf{t}) = \text{constante x } \prod_{i=1}^{p} \left\{ \beta e^{\alpha \boldsymbol{\tau}_i} \exp\left[-\left(\frac{\beta}{\alpha}\right)(e^{\alpha \boldsymbol{\tau}_i} - 1) \right] \right\}^{d_i} \left\{ \exp\left[-\left(\frac{\beta}{\alpha}\right)(e^{\alpha \boldsymbol{\tau}_i} - 1) \right] \right\}^{s_i} \tag{A.5}$$

e o logarítmo da verossimilhança é dado por,

$$l(t) = \text{constante} + \sum_{i=1}^{p} \left\{ d_i \left[\log k + \alpha \tau_i - \left(\frac{\beta}{\alpha}\right)(e^{\alpha \boldsymbol{\tau}_i} - 1) \right] - \left(\frac{\beta}{\alpha}\right) s_i (e^{\alpha \boldsymbol{\tau}_i} - 1) \right\}. \tag{A.6}$$

Definindo,

$$\begin{cases} T = \sum_{i=1}^{p} d_i \tau_i, \ D = \sum_{i=1}^{p} d_i \\ Q(\alpha) = s_i (e^{\alpha \boldsymbol{\tau}_i} - 1) + d_i (e^{\alpha \boldsymbol{\tau}_i} - 1) \end{cases} \tag{A.7}$$

o modelo pode ser reescrito como,

$$l(t) = cons\tan te + D \log K - \alpha T - \left(\frac{\beta}{\alpha}\right) Q(\alpha). \tag{A.8}$$

Apêndice B

O Modelo de Riscos Proporcionais de Cox

Em 1972 Cox propôs um modelo que permite a análise de dados de sobrevivência considerando as covariáveis de interesse para cada indivíduo. Cox assim como outros autores propõe modelar dados de sobrevivência na presença de covariáveis , por meio da função de risco.

Assim, a função de risco do i-ésimo indivíduo é dada por,

$$h_i\left(t|x_i\right) = h_0\left(t\right)\exp\left(\boldsymbol{\alpha}^{'}\mathbf{x}_i\right), \tag{B.1}$$

onde $h_0\left(t\right)$ é a função de risco base, α é o vetor de dimensão p de coeficientes de regressão desconhecidos e $\mathbf{x}_i$ é o vetor de dimensão p de covariáveis observadas para o i-ésimo indivíduo.

O modelo de Cox (1.9) é conhecido como sendo semi-paramétrico por assumir que as covariáveis atuam de forma multiplicativa no risco e por considerar $h_0\left(t\right)$ arbitrária, ou seja, por não ser assumida nenhuma forma paramétrica para $h_0\left(t\right)$.

As funções de taxa base acumulada bem como a correspondente função de sobrevivência são também de interesse e estas relacionam-se com a função de risco base por,

$$H_0\left(t\right) = \int_0^t h_0\left(t\right)dt \tag{B.2}$$

e ainda,

$$S\left(t|x\right) = \exp\left[-\int_0^t h\left(t|x\right) dt\right] \tag{B.3}$$

$$= S_0\left(t\right)^{\exp\left(\boldsymbol{\alpha}'\mathbf{x}\right)}, \tag{B.4}$$

onde $S_0\left(t\right) = \exp\left[-\int h_0\left(u\right) du\right]$.

B.1 A Função de Verossimilhança Parcial de Cox

O modelo de Cox caracteriza-se pelos coeficientes α que devem ser estimados a partir das observações amostrais. A presença do componente não-paramétrico inviabiliza, contudo, o uso do método da máxima verossimilhança. Para a estimação desses α Cox (1975) apresentou então a função de verossimilhança parcial.

Este modelo é conhecido como semi-paramétrico por assumir que as covariáveis atuam multiplicativamente no risco pela relação $g\left(x,\alpha\right) = \exp\left(\boldsymbol{\alpha}'\mathbf{x}_i\right)$ e por considerar $h_0\left(t\right)$ arbitrária, ou seja, por não ser assumida nenhuma forma paramétrica para $h_0\left(t\right)$. Sua denominação, como sendo de riscos proporcionais, se deve ao fato de a razão entre as funções de risco de dois indivíduos,

$$\frac{h\left(t|x_i\right)}{h\left(t|x_j\right)} = \frac{h_0\left(t\right)\exp\left(\alpha'\mathbf{x}_i\right)}{h_0\left(t\right)\exp\left(\boldsymbol{\alpha}'\mathbf{x}_j\right)} = \exp\left\{\boldsymbol{\alpha}'\left(x_i - x_j\right)\right\} \tag{B.5}$$

$(i, j = 1, ..., n \text{ e } i \neq j)$ não depender de t.

Observe que a constante α_0 não aparece no componente paramétrico $\exp\left(\boldsymbol{\alpha}'\mathbf{x}_i\right)$. Isto ocorre devido à presença do componente não-paramétrico que absorve este termo constante.

A função de verossimilhança para n indivíduos considerando dados censurados é estabelecida por,

$$L\left(t\right) = \prod_{i=1}^{n} f\left(t_i|x_i\right)^{\delta_i} S\left(t_i|x_i\right)^{1-\delta_i}. \tag{B.6}$$

Considerando $S_0\left(t\right) = \exp\left[-\int h_0\left(u\right) du\right]$ e de(1.7) temos,

$$L(\alpha) = \prod_{i=1}^{n} \left[h\left(t_i|x_i\right)\right]^{\delta_i} = \prod_{i=1}^{n} \left[h_0(t)\exp(\boldsymbol{\alpha}'\mathbf{x}_i)\right]^{\delta_i}. \tag{B.7}$$

Assumindo que $U(t)$ denota o conjunto de indivíduos que sobreviveram imediatamente antes do tempo t a função de verossimilhança parcial é dada por,

$$L(\alpha) = \prod_{i=1}^{n} \left[\frac{\exp\left(\mathbf{x}_i\boldsymbol{\alpha}'\right)}{\sum_{l\in U(t_i)}\exp\left(\mathbf{x}_i\boldsymbol{\alpha}'\right)}\right]^{\delta_i}. \tag{B.8}$$